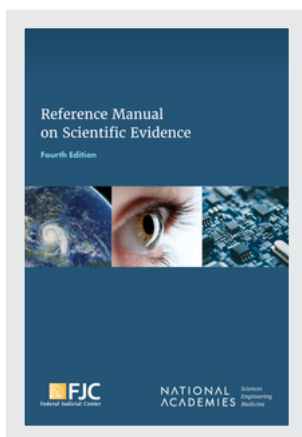


This PDF is available at <https://nap.nationalacademies.org/26919>



Reference Manual on Scientific Evidence: Fourth Edition (2025)

DETAILS

1682 pages | 6 x 9 | PAPERBACK

ISBN 978-0-309-70123-5 | DOI 10.17226/26919

CONTRIBUTORS

Committee on Science for Judges—Development of the Reference Manual on Scientific Evidence, Fourth Edition; Committee on Science, Technology, and Law; Policy and Global Affairs; National Academies of Sciences, Engineering, and Medicine; Federal Judicial Center

SUGGESTED CITATION

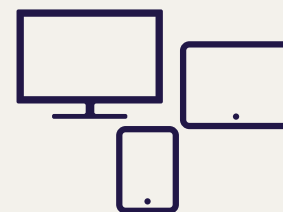
National Academies of Sciences, Engineering, and Medicine. 2025. *Reference Manual on Scientific Evidence: Fourth Edition*. Washington, DC: The National Academies Press. <https://doi.org/10.17226/26919>.

BUY THIS BOOK

FIND RELATED TITLES

Visit the National Academies Press at nap.edu and login or register to get:

- Access to free PDF downloads of thousands of publications
- 10% off the price of print publications
- Email or social media notifications of new titles related to your interests
- Special offers and discounts



All downloadable National Academies titles are free to be used for personal and/or non-commercial academic use. Users may also freely post links to our titles on this website; non-commercial academic users are encouraged to link to the version on this website rather than distribute a downloaded PDF to ensure that all users are accessing the latest authoritative version of the work. All other uses require written permission. ([Request Permission](#))

This PDF is protected by copyright and owned by the National Academy of Sciences; unless otherwise indicated, the National Academy of Sciences retains copyright to all materials in this PDF with all rights reserved.

Reference Manual on Scientific Evidence

Fourth Edition

Committee on Science for Judges—
Development of the Reference Manual
on Scientific Evidence, Fourth Edition

Committee on Science, Technology,
and Law

Policy and Global Affairs

NATIONAL ACADEMIES PRESS 500 Fifth Street, NW Washington, DC 20001

This project is funded in part by the Gordon and Betty Moore Foundation through Grant GBMF10224, and the National Science Foundation under Grant GBMF10224 to the National Academy of Sciences. Any opinions, findings, conclusions, or recommendations expressed in this publication do not necessarily reflect the views of any organization or agency that provided support for the project.

The Federal Judicial Center contributed to this publication in furtherance of the Center's statutory mission to conduct research and continuing education programs for judicial branch employees for the improvement of judicial administration. This manual is provided as a resource to assist judges in understanding relevant scientific methods and principles that may arise in their cases. The views expressed are those of the authors and not necessarily those of the Federal Judicial Center and its Board.

International Standard Book Number-13: 978-0-309-70123-5
Digital Object Identifier: <https://doi.org/10.17226/26919>
Library of Congress Control Number: 2025934845

This publication is available from the National Academies Press, 500 Fifth Street, NW, Keck 360, Washington, DC 20001; (800) 624-6242;
<https://nap.nationalacademies.org>.

The manufacturer's authorized representative in the European Union for product safety is Authorised Rep Compliance Ltd., Ground Floor, 71 Lower Baggot Street, Dublin D02 P593 Ireland; www.arccompliance.com.

Copyright 2025 by the National Academy of Sciences, National Academies of Sciences, Engineering, and Medicine and National Academies Press and the graphical logos for each are all trademarks of the National Academy of Sciences. All rights reserved.

Printed in the United States of America.

Suggested citation: National Academies of Sciences, Engineering, and Medicine and Federal Judicial Center. 2025. *Reference Manual on Scientific Evidence: Fourth Edition*. Washington, DC: National Academies Press. <https://doi.org/10.17226/26919>.

NATIONAL ACADEMIES OF SCIENCES, ENGINEERING, AND MEDICINE

The **National Academy of Sciences** was established in 1863 by an Act of Congress, signed by President Lincoln, as a private, nongovernmental institution to advise the nation on issues related to science and technology. Members are elected by their peers for outstanding contributions to research. Dr. Marcia McNutt is president.

The **National Academy of Engineering** was established in 1964 under the charter of the National Academy of Sciences to bring the practices of engineering to advising the nation. Members are elected by their peers for extraordinary contributions to engineering. Dr. Tsu-Jae Liu is president.

The **National Academy of Medicine** (formerly the Institute of Medicine) was established in 1970 under the charter of the National Academy of Sciences to advise the nation on medical and health issues. Members are elected by their peers for distinguished contributions to medicine and health. Dr. Victor J. Dzau is president.

The three Academies work together as the **National Academies of Sciences, Engineering, and Medicine** to provide independent, objective analysis and advice to the nation and conduct other activities to solve complex problems and inform public policy decisions. The National Academies also encourage education and research, recognize outstanding contributions to knowledge, and increase public understanding in matters of science, engineering, and medicine.

Learn more about the National Academies of Sciences, Engineering, and Medicine at **www.nationalacademies.org**.

FEDERAL JUDICIAL CENTER

Congress established the Federal Judicial Center to support the efficient and effective administration of the federal courts through research and continuing education of judges and court employees. Its separate status within the judicial branch, its specific missions, and its specialized expertise enable it to pursue and encourage critical and careful examination of ways to improve judicial administration. The Federal Judicial Center has no policymaking or enforcement authority; its role is to provide accurate, objective information and education and to encourage thorough and candid analysis of policies, practices, and procedures.

By statute, the Chief Justice of the United States chairs the Center's Board, which also includes the director of the Administrative Office of the U.S. Courts and seven judges elected by the Judicial Conference of the United States. The Board appoints the Center's director and deputy director; the director appoints the Center's staff. Since its founding in 1967, the Center has had twelve directors. John S. Cooke was director from September 2018–August 2025; Judge Robin L. Rosenberg became director in August 2025. Clara Altman has been deputy director since 2018.

The organization of the Center reflects its primary statutory mandates. The Director's Office is responsible for the Center's overall management and its relations with other organizations. The Research Division examines and evaluates current and alternative federal court practices and policies. This research assists the Judicial Conference of the United States. The Center's research also contributes to its education mission. The Education Division plans and produces education and training for judges and court staff, including in-person programs, video programs, publications, and Web-based programs and resources. The Federal Judicial History Office helps courts and others study and preserve federal judicial history. The International Office provides information to judicial and legal officials from foreign countries and informs federal judicial personnel of developments in international law and other court systems that may affect their work. Two units of the Director's Office—the Information Technology Office and the Editorial & Information Services Office—support Center missions through technology, editorial, and design assistance, and organization and dissemination of Center resources.

For more information, see **www.fjc.gov**.

*Committee on Science for Judges—Development of the
Reference Manual on Scientific Evidence, Fourth Edition*

NANCY D. FREUDENTHAL (*Co-Chair*), Judge, U.S. District Court for the District of Wyoming

FRED H. GAGE (NAS, NAM) (*Co-Chair*), Adler Professor, Laboratory of Genetics, The Salk Institute for Biological Studies

RUSS BIAGIO ALTMAN, Kenneth Fong Professor of Bioengineering, Genetics, Medicine, Biomedical Data Science, and (by courtesy) Computer Science, Stanford University

DAVID G. CAMPBELL, Senior Judge, U.S. District Court for the District of Arizona

ALICIA L. CARRIQUIRY (NAM), Distinguished Professor of Liberal Arts and Sciences and the President's Professor of Statistics, Iowa State University

LYNN R. GOLDMAN (NAM), Michael and Lori Milken Dean and Professor of Environmental and Occupational Health, Milken Institute School of Public Health, George Washington University

BRIAN W. KERNIGHAN (NAE), William O. Baker '39 Professor, Department of Computer Science, Princeton University

PRAMOD P. KHARGONEKAR, Vice Chancellor for Research and Distinguished Professor of Electrical Engineering and Computer Science, University of California, Irvine

GOODWIN LIU, Associate Justice, Supreme Court of California

SHOBITA PARTHASARATHY, Professor of Public Policy and Women's and Gender Studies, Ford School of Public Policy and Director, Science, Technology, and Public Policy Program, University of Michigan

PATTI B. SARIS, Judge, U.S. District Court for the District of Massachusetts

THOMAS D. SCHROEDER, Judge, U.S. District Court for the Middle District of North Carolina

DAVID S. TATEL, Judge (retired), U.S. Court of Appeals for the District of Columbia Circuit

National Academies Staff

ANNE-MARIE MAZZA, Project Director and Senior Director, Committee on Science, Technology, and Law, National Academies of Sciences, Engineering, and Medicine

STEVEN KENDALL, Senior Program Officer, Committee on Science, Technology, and Law, National Academies of Sciences, Engineering, and Medicine

RENEE DALY, Senior Program Assistant, Committee on Science, Technology, and Law, National Academies of Sciences, Engineering, and Medicine (until April 2025)

DOMINIC LOBUGLIO, Senior Program Assistant, Committee on Science, Technology, and Law, National Academies of Sciences, Engineering, and Medicine (until February 2023)

MATTHEW COWAN, Christine Mirzayan Science and Technology Policy Graduate Fellow (until May 2023)

Federal Judicial Center Staff

ELIZABETH C. WIGGINS, Director, Research Division, Federal Judicial Center

JASON A. CANTONE, Senior Research Associate, Federal Judicial Center

MEGHAN A. DUNN, Senior Research Associate, Federal Judicial Center

ELLEN URHEIM, AAAS Science & Technology

Policy Fellow, Federal Judicial Center (until August 2025); Contractor with Federal Judicial Center (since September 2025)

REBEKAH PETROFF, AAAS Science & Technology Policy Fellow, Federal Judicial Center (until August 2024)

ALEXIS ALLEGRA, AAAS Science & Technology Policy Fellow, Federal Judicial Center (until August 2023)

RESHMINA WILLIAM, AAAS Science & Technology Policy Fellow, Federal Judicial Center (until August 2022)

Consultant

JOE S. CECIL (until June 2024)

Board of the Federal Judicial Center

THE CHIEF JUSTICE OF THE UNITED STATES (*Chair*)

KATHLEEN CARDONE, Judge, U.S. District Court for the Western District of Texas

R. GUY COLE, JR., Judge, U.S. Court of Appeals for the Sixth Circuit

SARA L. ELLIS, Judge, U.S. District Court for the Northern District of Illinois

RALPH R. ERICKSON, Judge, U.S. Court of Appeals for the Eighth Circuit

MICHELLE M. HARNER, Judge, U.S. Bankruptcy Court for the District of Maryland

SUZANNE MITCHELL, Magistrate Judge, U.S. District Court for the Western District of Oklahoma

LYNN WINMILL, Judge, U.S. District Court for the District of Idaho

ROBERT J. CONRAD, JR., Judge and Director of the Administrative Office of the U.S. Courts

Committee on Science, Technology, and Law

MARTHA MINOW (*Co-Chair*), Professor of Law, Harvard Law School
HAROLD VARMUS (NAS/NAM) (*Co-Chair*), Lewis Thomas University

Professor, Meyer Cancer Center, Weill Cornell Medicine

DAVID APATOFF, Senior Partner, Arnold & Porter

ERWIN CHEMERINSKY, Dean and Jesse H. Choper Distinguished
Professor of Law, University of California, Berkeley School of Law

ELLEN WRIGHT CLAYTON (NAM), Professor of Law, Professor of
Health Policy, Vanderbilt University Medical Center

JOHN S. COOKE, Director, Federal Judicial Center (until August 2025)

JENNIFER EBERHARDT (NAS), Professor of Psychology, Stanford
University

KENNETH C. FRAZIER, Chairman, Health Assurance Initiatives,
General Catalyst

CAROL GREIDER (NAS/NAM), Distinguished Professor of Molecular,
Cell, and Developmental Biology, University of California, Santa Cruz

STEVE HYMAN (NAM), Harald McPike Professor of Stem Cell and
Regenerative Biology, Core Institute Member and Director, Broad
Institute of MIT and Harvard

JON M. KLEINBERG (NAS/NAE), Tisch University Professor, Cornell
University

BARBARA MCGAREY, Deputy General Counsel (retired),
U.S. Department of Health and Human Services

ERNEST J. MONIZ, Cecil and Ida Green Professor of Physics and
Engineering Systems, Post-Tenure, Massachusetts Institute of
Technology

KIMANI PAUL-EMILE, Professor of Law, Fordham University
Law School

K. SABEEL RAHMAN, Professor of Law, Cornell Law School

NATALIE RAM, Professor of Law, University of Maryland Francis King
Carey School of Law

JULIE ROBINSON, Senior Judge, U.S. District Court for the District of
Kansas

PATTI B. SARIS, Judge, U.S. District Court for the District of Massachusetts

VICKI SATO, Chair, VIRbio

BARBARA SCHAAL (NAS), Mary-Dell Chilton Distinguished Professor,
Washington University in St. Louis

JOSHUA SHARFSTEIN, Vice Dean and Professor, Johns Hopkins
Bloomberg School of Public Health

CLIFFORD TABIN (NAS), Leder Professor and Chair, Harvard Medical
School

Staff

ANNE-MARIE MAZZA, Senior Director

STEVEN KENDALL, Senior Program Officer

RENEE DALY, Senior Program Assistant (until April 2025)

Reviewers

This *Reference Manual on Scientific Evidence* was reviewed in draft form by individuals chosen for their diverse perspectives and technical expertise. The purpose of this independent review is to provide comments that will assist the National Academies of Sciences, Engineering, and Medicine in making each publication as sound as possible and to ensure that it meets the institutional standards for quality, objectivity, evidence, and responsiveness to the statement of task. The review comments and draft manuscripts remain confidential to protect the integrity of the deliberative process.

We thank the following individuals for their review of this manual:

HELEN ADAMS, U.S. District Court for the Southern District of Iowa

BRUCE ALBERTS (NAS), University of California, San Francisco

PAOLA ARLOTTA (NAS, NAM), Harvard University

ORLEY ASHENFELTER (NAS), Princeton University

PAUL BARBADORO, U.S. District Court for the District of New Hampshire

JOHN BUTLER, U.S. Department of Commerce

CLEOPATRA CABUZ (NAE), Honeywell Industrial Safety (retired)

FRED CATE, Indiana University

VINTON CERF (NAS, NAE), Google

EDWARD CHENG, Vanderbilt University

ELLEN WRIGHT CLAYTON (NAM), Vanderbilt University

CARL CRANOR, University of California, Riverside

CHRIS FIELD (NAS), Stanford University

ELIZABETH GEDDES, Shihata & Geddes

CHARLES HAAS (NAE), Drexel University

SARA HUSTON, Ann & Robert H. Lurie Children's Hospital of Chicago

STEVEN HYMAN (NAM), Broad Institute of MIT and Harvard

RAYMOND JACKSON, U.S. District Court for the Eastern District of Virginia

KATHLEEN HALL JAMIESON (NAS), University of Pennsylvania

PAUL JANICKE, University of Houston

KAREN KAFADAR, University of Virginia

MARGARET BULL KOVERA, John Jay College of Criminal Justice

ERIC LANDER (NAS, NAM), Broad Institute of MIT and Harvard

JULIA LEIGHTON, Public Defender Service for the District of Columbia (retired)

PATRICK MALONE, Patrick Malone & Associates

RICHARD A. MESERVE (NAE), Carnegie Science

ROBERT MILLER, U.S. District Court for the Northern District of Indiana

DAVID OWEN, University of South Carolina (retired)

MARIA POLYAKOVA, Stanford University

ANTHONY PORCELLI, U.S. District Court for the Middle District of
Florida

WILLIAM PRESS (NAS), The University of Texas at Austin

JED RAKOFF, U.S. District Court for the Southern District of New York

NATALIE RAM, University of Maryland

MARGARET RILEY, University of Virginia

VICTORIA ROBERTS, U.S. District Court for the Eastern District of
Michigan (retired)

JULIE ROBINSON, U.S. District Court for the District of Kansas

BARBARA ROTHSTEIN, U.S. District Court for the Western District of
Washington

JOELLEN RUSSELL, University of Arizona

JOHN SAMET (NAM), University of Colorado

NATHAN A. SCHACHTMAN, UB Greensfelder

FRANCIS SHEN, Harvard University

MICHAEL SIMON, U.S. District Court for the District of Oregon

WILLIAM SMITH, U.S. District Court for the District of Rhode Island

HAROLD SOX (NAM), Dartmouth University (retired)

HOLDEN THORP, American Association for the Advancement of Science

WENDY WAGNER, The University of Texas at Austin

LYNN WINMILL, U.S. District Court for the District of Idaho

JOHN WIXTED, University of California, San Diego

DIANE WOOD, University of Chicago

DONALD WUEBBLES, University of Illinois

JAMES WYNN, U.S. Court of Appeals for the Fourth Circuit

Although the reviewers listed above provided many constructive comments and suggestions, they were not asked to endorse the *Reference Manual on Scientific Evidence*, nor did they see the final draft before its release. The review of this edition was overseen by Judge Jed S. Rakoff, U.S. District Court for the Southern District of New York, and Judge Kathleen M. O'Malley (retired), Kate O'Malley LLC. They were responsible for making certain that an independent examination of the manual was carried out in accordance with the standards of the National Academies and that all review comments were carefully considered. Responsibility for the final content rests with the Committee on Science for Judges—Development of the Reference Manual on Scientific Evidence, Fourth Edition, the National Academies, and the Federal Judicial Center.

Foreword

I am pleased to introduce this fourth edition of the *Reference Manual on Scientific Evidence*. A joint product of the Federal Judicial Center and the National Academies of Sciences, Engineering, and Medicine, the manual has for three decades assisted judges in considering all manner of statistical and scientific evidence offered in litigation. In so doing, it has helped bring about better and fairer legal decisions.

It will surprise no one that science and law are today often intertwined. We live in an era of science and technology, with legal controversies reflecting that fact. And no sooner does one scientific field begin to become accessible to judges than another one emerges or fundamentally evolves. In the coming years, judges will confront lawsuits relating, for example, to artificial intelligence, climate science, and epidemiology. Even disputes not about science may involve statistical, survey, or other technical evidence of novel kinds.

Enter this manual. Judges typically are generalists, and they often lack extensive background in the sciences. They can learn, as they do on many subjects, through the adversary process, applying their critical faculties to the claims of parties. But sometimes it also helps to have a dispassionate guide. The manual is the product of close collaboration among highly respected scientists, engineers, judges, and lawyers. It delves into the scientific subjects that judges most often face. It explains scientific approaches and explores scientific uncertainties and limits. It aids in assessing the uses—and the misuses—of scientific and other technical evidence. The manual will hardly resolve every scientific issue arising in the law; that is not, and cannot be, its ambition. Yet case in and case out, the instruction that the manual offers in scientific principles and methods can improve the quality of judicial decision making.

In this way, the *Reference Manual on Scientific Evidence* exemplifies the benefits that can accrue from cooperation between those proficient in legal analysis and those expert in technical subjects. Judges, along with the lawyers who argue before them, will do their jobs better when they recognize what they can learn from the scientific community. And the law will become stronger as it further reflects sound science.

ELENA KAGAN
Associate Justice
United States Supreme Court

Summary Table of Contents

A detailed Table of Contents appears at the front of each reference guide.

Preface, xvii

The Admissibility of Expert Testimony, 1

Liesa L. Richter & Daniel J. Capra

How Science Works, 47

Michael Weisberg & Anastasia Thanukos

Reference Guide on Forensic Feature Comparison Evidence, 113

Valena E. Beety, Jane Campbell Moriarty, & Andrea L. Roth

Reference Guide on Human DNA Identification Evidence, 207

David H. Kaye

Reference Guide on Eyewitness Identification, 361

Thomas D. Albright & Brandon L. Garrett

Reference Guide on Statistics and Research Methods, 463

David H. Kaye & Hal S. Stern

Reference Guide on Multiple Regression and Advanced Statistical Models, 577

Daniel L. Rubinfeld & David Card

Reference Guide on Survey Research, 681

Shari Seidman Diamond, Matthew Kugler, & James N. Druckman

Reference Guide on Estimation of Economic Damages, 749

Mark A. Allen, Carlos Brain, & Filipe Lacerda

Prologue to the *Reference Guide on Exposure Science and Exposure Assessment*, the
Reference Guide on Epidemiology, and the *Reference Guide on Toxicology*, vii

Reference Guide on Exposure Science and Exposure Assessment, 831

M. Elizabeth Marder & Joseph V. Rodricks

Reference Guide on Epidemiology, 897

Steve C. Gold, Michael D. Green, Jonathan Chevrier, & Brenda Eskenazi

Reference Guide on Toxicology, 1027

David L. Eaton, Bernard D. Goldstein, & Mary Sue Henifin

Reference Guide on Medical Testimony, 1105

John B. Wong, Lawrence O. Gostin, & Oscar A. Cabrera

Reference Guide on Neuroscience, 1185

Henry T. Greely & Nita A. Farahany

Reference Manual on Scientific Evidence

Reference Guide on Mental Health Evidence, 1269	
Kirk Heilbrun, David DeMatteo, & Paul S. Appelbaum	
Reference Guide on Engineering, 1353	
Chaouki T. Abdallah, Bert Black, & Edl Schamiloglu	
Reference Guide on Computer Science, 1409	
Brian N. Levine, Joanne Pasquarelli, & Clay Shields	
Reference Guide on Artificial Intelligence, 1481	
James E. Baker & Laurie N. Hobart	
Reference Guide on Climate Science,* 1561	
Jessica Wentz & Radley Horton	
Appendix. Biographical Information of Committee and Staff, 1653	

*Following the December 31, 2025, publication of the Fourth Edition of *Reference Manual on Scientific Evidence*, on February 6, 2026, the Federal Judicial Center decided to omit the climate science chapter from their website.

Preface

This fourth edition of the *Reference Manual on Scientific Evidence* is the product of an ongoing collaboration between the Federal Judicial Center and the National Academies of Sciences, Engineering, and Medicine. The contributions of content by distinguished scholars from across disciplines have resulted in an impartial and reliable primary reference source on science for judges, which we set as our goal. We are grateful to the National Science Foundation and the Gordon and Betty Moore Foundation for the support that made this new edition possible.

Much has changed since the publication of the first edition of the *Reference Manual on Scientific Evidence* in 1994. Over the years, judges have become more comfortable in exercising a “gatekeeper” function over the admissibility of scientific evidence. Nevertheless, litigation is becoming ever more complex, and aggregate litigation now represents a larger share of the federal civil docket. Scientific, medical, and engineering evidence has also become more complex, and judges must often wrestle with unfamiliar and emerging science and technology. As a consequence, the rules governing admissibility have been revised to deal with emerging problems. For example, amendments to Rule 702 of the Federal Rules of Evidence were recently adopted to discourage testifying experts from overstating the value of evidence underlying such testimony. The dynamic and evolving presence of science in litigation necessitates engagement between the scientific and legal communities. The fourth edition of the *Reference Manual on Scientific Evidence* reflects this reality.

The primary (but not exclusive) audience for the manual is members of the judiciary who face challenging evidentiary issues related to science and technology. The manual is intended to assist judges in identifying issues commonly in dispute and to help judges reach an informed and reasoned assessment of those issues based on expert evidence that is faithful to the law and within the boundaries of scientifically sound knowledge. The manual is *not* intended to instruct judges concerning what evidence should be admissible or to establish minimum standards for acceptable scientific testimony.

In the new edition, all reference guides from the previous edition have undergone extensive revision or have been written by new authors. New guides have been added to address eyewitness identification, computer science, artificial intelligence, and climate science, along with a new foreword from Associate Justice of the U.S. Supreme Court Elena Kagan.

We are honored to have co-chaired the committee tasked with the development of this new edition of the *Reference Manual on Scientific Evidence*. Convened under the auspices of the National Academies’ Committee on Science, Technology, and Law, the committee comprised six judges and seven experts from across science and engineering. All have given generously of their time, judgment, and wisdom to produce a volume that provides a crucial resource for the legal and scientific communities.

We are especially grateful to the authors who worked so diligently in drafting manuscripts for the committee's review and who responded so thoughtfully to the committee's and external reviewers' comments.

We are also grateful for the input of federal judges who responded to a December 2020 survey about how they used the manual and any changes and additions they thought would make it more helpful and to the co-chairs of the National Academies' 2021 workshop entitled *Emerging Areas of Science, Engineering, and Medicine for the Courts*.¹ The survey responses, the workshop, and 2019 and 2020 discussions² on science and the courts convened by Dr. Marcia McNutt, president of the National Academy of Sciences, identified many topics and issues that appear in this new edition of the manual.

Additionally, the committee received substantial encouragement from John Cooke, former director of the Federal Judicial Center, and Alan Thompkins, former deputy division director, acting division director, and acting deputy assistant director in the Social, Behavioral and Economic Sciences Directorate. The committee benefitted greatly from the work of Federal Judicial Center staff, particularly Meghan Dunn and Jason Cantone, and from the staff of the National Academies, particularly Steven Kendall, Renee Daly, and Dominic LoBuglio, and from editors Sally-Anne Cleveland, Nathan Dotson, Susanna McCrea, Clare O'Shea, and particularly Geoffrey Erwin.

Special commendation goes to Joe Cecil, project consultant, and Anne-Marie Mazza, project director and senior director of the Committee on Science, Technology, and Law of the National Academies, along with Beth Wiggins, Research Division director for the Federal Judicial Center. Their many contributions in shepherding the manual to completion have been invaluable.

NANCY D. FREUDENTHAL
AND FRED H. GAGE
Committee Co-Chairs

1. See <https://www.nationalacademies.org/science-for-courts-workshop>.

2. Science in the Courts: Amicus Briefs and the Law, Feb. 2019; and Responsiveness of National Academies' Reports to the Needs of the Courts, Nov. 2020. These discussions were supported by the Ralph J. Cicerone Endowment Fund and the Annenberg Foundation Trust at Sunnyslands.

The Admissibility of Expert Testimony

LIESA L. RICHTER AND DANIEL J. CAPRA

Liesa L. Richter, J.D., is George Lynn Cross Research Professor, Floyd & Martha Norris Chair in Law, University of Oklahoma College of Law, and Academic Consultant to the Judicial Conference Advisory Committee on Evidence Rules.

Daniel J. Capra, J.D., is Philip Reed Professor of Law, Fordham Law School, and Reporter to the Judicial Conference Advisory Committee on Evidence Rules.

CONTENTS

Introduction,	3
Federal Rule of Evidence 702: An Overview,	3
History of Rule 702,	4
The Pre-Rules <i>Frye</i> Standard and Original Rule 702,	4
The Supreme Court’s <i>Daubert</i> Trilogy,	5
Amendments to Rule 702,	13
The 2000 Amendment to Rule 702,	14
Rule 702: A Qualified Expert,	14
Rule 702(a): The Expert’s Opinion “Will Help”	
the Trier of Fact,	15
Rule 702(b): The Expert’s Opinion Testimony Is Based	
on Sufficient Facts or Data,	16
Rule 702(c): The Testimony Is the Product of Reliable	
Principles and Methods,	18
Rule 702(d): The Witness Has Applied the Principles	
and Methods Reliably to the Facts of the Case,	19
The 2023 Amendment to Rule 702,	22
Procedures for Resolving Rule 702 Motions and Objections,	24
<i>Daubert</i> Motions and Summary Judgment,	25
Credibility Determinations and Rule 702 Motions,	26
Important Procedural Takeaways,	27
Recurring Issues in the Application of Rule 702,	28
Extrapolation and the “Analytical Gap,”	29
Weight of the Evidence Analysis: Closing the Analytical Gap,	31
Reliance on Anecdotal Evidence,	34
Reliance on Temporal Proximity,	35

Reference Manual on Scientific Evidence

Insufficient Connection Between the Expert's Opinion and
the Facts in the Case: The Requirement of "Fit," 36
Lack of Testing, 37
The Problem of Overstatement, 39
Experts Relying on Technical or Other Specialized Knowledge, 42
Conclusion, 45

Introduction

The trial process operates to resolve disputed facts and claims through the presentation of evidence. The vast majority of evidence presented at trial comes in through the testimony of witnesses. Many of these witnesses are lay witnesses with personal knowledge of the facts and underlying events relevant to the issues in dispute. Often, however, the finder of fact requires assistance from witnesses who possess scientific, technical, or other specialized knowledge. Such witnesses offer expert opinion testimony that, when properly validated, helps jurors and judges resolve complex issues that are outside common knowledge and experience. Federal Rule of Evidence 702 regulates the admissibility of expert opinion testimony. Pursuant to Rule 702, the federal trial judge serves as a gatekeeper for expert opinion testimony, ensuring that only reliable expert testimony is considered by the factfinder.

The enactment, in 1975, of Federal Rule of Evidence 702, the interpretation of that rule by the U.S. Supreme Court in *Daubert v. Merrell Dow Pharmaceuticals, Inc.*,¹ and the further amendments to Rule 702 as recently as 2023, have effected a gradual but significant change in how federal courts treat the admissibility of scientific evidence. Previously, a district court had to decide whether the proffered evidence was deduced from scientific principles generally accepted in the corresponding scientific community. Now, the district court must itself determine whether it is more likely than not that these principles are reliable and have been reliably applied to the facts of the case. This requires the court to become conversant with the science at issue—helped perhaps by this publication. At the same time, courts are cautioned not to prefer one valid scientific theory over another and to admit even differing expert opinions that satisfy the Rule 702 reliability standard. This reference guide examines how courts have tried to grapple with this difficult gatekeeper role.

Federal Rule of Evidence 702: An Overview

When a party objects to the admissibility of expert opinion testimony, Rule 702 requires a trial judge to analyze the basis of the challenge and to make certain findings before admitting the testimony.² Rule 104(a) of the Federal Rules of Evidence requires trial judges to decide preliminary questions regarding the admissibility of evidence by a preponderance of the evidence. Rule 104(a) applies to

1. 509 U.S. 595 (1993).

2. As with other evidence rules, a trial judge need not make findings under Rule 702 in the absence of an objection to expert opinion testimony. See Fed. R. Evid. 702 advisory committee's note to 2023 amendment.

the trial judge's findings under Rule 702, as it does to the findings under most of the evidence rules.³ To admit expert opinion testimony over objection, a trial judge must consider the following aspects of the testimony, if challenged. First, the judge must find the witness *qualified* to opine on the subject matter at issue. The judge must also find that the witness's scientific, technical, or other specialized knowledge *will help* the trier of fact. Before admitting the expert's opinion, the court must find that the opinion is based on *sufficient facts or data*. In addition, the court must determine that the witness's opinion is the product of *reliable principles and methods* and that the opinion reflects a *reliable application* of those principles and methods to the facts of the case. And a 2023 amendment to Rule 702 clarifies that the trial judge should determine that the expert witness does not overstate the conclusions that may be drawn from the methodology. The proponent of the testimony has the burden of satisfying the challenged aspects of Rule 702 by a preponderance of the evidence. Although Rule 702 characterizes opinion testimony based on scientific, technical, or other specialized knowledge as "expert" opinion testimony, some federal opinions caution trial courts to refrain from referring to witnesses as "experts" to avoid giving them undue influence with the factfinder.⁴

History of Rule 702

The Pre-Rules *Frye* Standard and Original Rule 702

In order to understand the current Rule 702 standard, it is helpful to be familiar with the history of expert opinion testimony in American trials and the standard

3. The Supreme Court, in *Bourjaily v. United States*, 483 U.S. 171 (1987), held that Rule 104(a) requires preliminary questions regarding the admissibility of evidence to be decided by the trial judge using a preponderance of the evidence standard.

4. See, e.g., *United States v. Maya*, 966 F.3d 493, 505 (6th Cir. 2020) ("We have cautioned district courts not to declare a witness an expert in front of the jury."); *United States v. Bartley*, 855 F.2d 547, 552 (8th Cir. 1988) (noting that "[s]uch an offer and finding by the Court might influence the jury in its evaluation of the expert and the better procedure is to avoid an acknowledgment of the witnesses' expertise by the Court"). See also American Bar Association Civil Trial Practice Standard 14 (2007) ("The court should not, in the presence of the jury, declare that a witness is qualified as an expert or to render an expert opinion, and counsel should not ask the court to do so" because "use of the term 'expert' may appear to a jury to be a kind of judicial imprimatur that favors the witness."); Fed. R. Evid. 702 advisory committee's note to 2000 amendment ("[T]here is much to be said for a practice that prohibits the use of the term 'expert' by both the parties and the court at trial. Such a practice 'ensures that trial courts do not inadvertently put their stamp of authority' on a witness' opinion, and protects against the jury's being 'overwhelmed by the so-called "experts."'") (quoting Hon. Charles R. Richey, *Proposals to Eliminate the Prejudicial Effect of the Use of the Word "Expert" Under the Federal Rules of Evidence in Criminal and Civil Jury Trials*, 154 F.R.D. 537, 559 (1994)).

The Admissibility of Expert Testimony

of admissibility that applied before the Federal Rules. In 1923, in the case of *Frye v. United States*, the D.C. Circuit Court of Appeals held that polygraph evidence was properly excluded because it had not gained “general acceptance” in the scientific community.⁵ Thereafter, the *Frye* “general acceptance” standard was widely adopted by federal and state courts alike to regulate the admissibility of expert testimony.⁶ When the Federal Rules of Evidence took effect in 1975, the *Frye* “general acceptance” standard reigned. As originally enacted, Rule 702 provided simply that a qualified witness could testify if scientific, technical, or other specialized knowledge would “assist” the trier of fact to understand the evidence or to determine a fact in issue. It did not reference the *Frye* standard or give any clue to its continued application under the rules.

The Supreme Court’s *Daubert* Trilogy

Until 1993, there was considerable controversy over whether Federal Rule of Evidence 702 incorporated the *Frye* “general acceptance” test. Then, in *Daubert v. Merrell Dow Pharmaceuticals, Inc.*,⁷ the Supreme Court unanimously rejected the *Frye* “general acceptance” standard as the exclusive test for assessing the admissibility of scientific expert testimony under Rule 702.

Daubert was a case in which the plaintiffs claimed that the use of anti-nausea drug Bendectin during pregnancy caused birth defects in their babies. The district court granted summary judgment for the defendant after excluding the plaintiffs’ expert on causation under *Frye*’s general acceptance standard. Because the expert’s reliance on animal studies, chemical structure analysis, and reanalysis of epidemiological studies had not been generally accepted as methods for determining causation, the trial court excluded the expert’s opinion testimony. On appeal, the Supreme Court scrutinized the language of Rule 702 and found nothing pointing to the continued application of the “general acceptance” standard. The Court further opined that a rigid “general acceptance” standard for the admissibility of scientific evidence would be at odds with the “liberal thrust” of the Federal Rules of Evidence and their otherwise flexible approach to opinion testimony. Therefore, the Court held that Rule 702 did not adopt the *Frye* “general acceptance” standard as the sole measure of the admissibility of expert opinion testimony.

5. 293 F. 1013 (D.C. Cir. 1923).

6. See *Daubert v. Merrell Dow Pharms., Inc.*, 509 U.S. 579, 585 (1993) (“In the 70 years since its formulation in the *Frye* case, the ‘general acceptance’ test has been the dominant standard for determining the admissibility of novel scientific evidence at trial.”). For a discussion of the admissibility of scientific evidence under the *Frye* standard, see Edward Imwinkelried et al., *Courtroom Criminal Evidence*, Vol. 1 §§ 608–613 (7th ed. 2022).

7. 509 U.S. 579 (1993).

Although it rejected “general acceptance” as the sine qua non of admissibility, the Court found limits on the admissibility of scientific expert opinion testimony implied by the terms “scientific” and “knowledge” in Rule 702. According to *Daubert*, a trial judge must ensure that scientific evidence is not only relevant, but *reliable*. Rule 702 requires the trial judge to act as a “gatekeeper” for scientific expert testimony, to ensure that it truly proceeds from “scientific . . . knowledge.” The Court stated that an inference or assertion must be derived from the scientific method in order to qualify as scientific knowledge.

Essentially, the Court held that a trial judge must evaluate proffered scientific expert testimony to ensure that it is more likely than not reliable; concerns about the validity and reliability of expert testimony cannot simply be referred to the jury as questions of weight. The Court made this clear by noting that Rule 104(a) provides the standard for admitting challenged expert testimony. Rule 104(a) had previously been interpreted by the Court as requiring admissibility requirements to be established by a preponderance of the evidence.⁸ Under *Daubert*, therefore, the judge must find it more likely than not that the expert’s methods are reliable and that they are reliably applied to the facts at hand.

The move from *Frye* to *Daubert* dramatically changed the role of federal district judges. Under *Frye*, the trial judge could essentially count heads in the relevant scientific community to determine whether a particular methodology was generally accepted. Under *Daubert*, however, the decision about reliability cannot be completely outsourced to the scientific community; it is the trial judge who must examine the proffered scientific opinion testimony to determine whether it is reliable. Thus, a district judge faces the challenging task of understanding the science behind a proffered expert opinion sufficiently to assess its reliability. Chief Justice Rehnquist, dissenting from the majority’s gatekeeper holding in *Daubert*, lamented that the majority had imposed on trial judges “either the obligation or the authority to become amateur scientists in order to perform that [gatekeeper] role.”⁹

To assist judges in this new gatekeeping task, the majority in *Daubert* set forth a five-factor, nondispositive, nonexclusive, “flexible” test to be employed by the trial court under Rule 702 in determining the “validity” of scientific evidence. These factors—the famous “*Daubert* factors”—are:

1. whether the technique or theory can be or has been tested;
2. whether the theory or technique has been subject to peer review and publication;

8. The Court held that questions of admissibility under Rule 104(a) must be determined by a preponderance of the evidence in *Bourjaily v. United States*, 483 U.S. 171 (1987).

9. 509 U.S. at 601.

The Admissibility of Expert Testimony

3. the known or potential rate of error;
4. the existence and maintenance of standards and controls; and
5. the degree to which the theory or technique has been generally accepted in the scientific community.

The factors described by the Court demand an objective assessment of an expert's technique or theory. With respect to testing, the Court explained that whether a scientific technique or theory has been or can be tested is ordinarily a key question in assessing its reliability because scientific methodology "is based on generating hypotheses and testing them to see if they can be falsified." The Court explained that publication is an important consideration because submission to the scrutiny of the scientific community increases the likelihood that substantive flaws in the methodology will be detected. Still, the Court recognized that publication is not an absolute requirement of admissibility given that well-grounded but innovative theories may not have been published. Finally, the Court explained that the general acceptance of a technique remains relevant to, although not dispositive of, admissibility. The Court explained that widespread acceptance can be an important factor in finding a methodology reliable.

The Ninth Circuit applied the gatekeeping standard on remand in the *Daubert* case, offering additional insights into the newly articulated test.¹⁰ The court held that summary judgment was properly granted against the plaintiffs because their experts' testimony that Bendectin caused limb reduction was inadmissible under Rule 702. In so holding, the Ninth Circuit focused on "whether the experts are proposing to testify about matters growing naturally and directly out of research they have conducted independent of the litigation, or whether they have developed their opinions expressly for purposes of testifying." According to the court, "a scientist's normal workplace is the lab or the field, not the courtroom or the lawyer's office."¹¹ Thus, when expert opinion testimony is based on research conducted independent of litigation, this "provides important, objective proof that the research comports with the dictates of good science." The court reasoned that "independent research carries its own indicia of reliability, as it is conducted, so to speak, in the usual course of business and must normally satisfy a variety of standards to attract funding and institutional support."¹²

10. *Daubert v. Merrell Dow Pharms., Inc.*, 43 F.3d 1311, 1317–19 (9th Cir. 1995).

11. *Id.*

12. *Id.* In casting doubt on research conducted in anticipation of litigation, the *Daubert on remand* court referred to research on issues of *general*, as opposed to specific, causation. In a toxic tort case, the plaintiff must show both that the toxic substance causes the harm suffered by the plaintiff (the issue of general causation) and that the toxic substance actually did cause the plaintiff's harm in the instant case (the issue of specific causation). *Daubert on remand* expressed concern about research into general causation that is conducted in anticipation of litigation. But an expert cannot be expected to conduct research on specific causation—the cause of a particular plaintiff's

The Ninth Circuit did not hold that an opinion that is *not* based on independent research is inadmissible. Instead, the court explained that “the party proffering it must come forward with other objective, verifiable evidence that the testimony is based on scientifically valid principles” if the expert’s testimony is not based on independent research.¹³ According to the Ninth Circuit, this means that “the experts must explain precisely how they went about reaching their conclusions and point to some objective source—a learned treatise, the policy statement of a professional association, a published article in a reputable scientific journal or the like—to show that they have followed the scientific method, as it is practiced by (at least) a recognized minority of scientists in their field.”¹⁴

Applying these principles to the testimony of the plaintiffs’ experts, the *Daubert on remand* court found that the *Daubert* standards had not been met. The experts’ research on the question of general causation was prepared solely for purposes of Bendectin litigation. It had not been peer reviewed, nor had it even been deemed worthy of comment by the scientific community. Finally, the experts had not explained their methodology or verified it by reference to objective sources. The court concluded that it had been presented “with only the experts’ qualifications, their conclusions and their assurances of reliability. Under *Daubert*, that’s not enough.”¹⁵

Federal opinions since *Daubert* have continued to articulate factors that may be important in assessing the admissibility of expert testimony, which are well summarized by the court in *In re Marriott Int’l, Inc., Customer Data Sec. Breach Litig.*: whether the expert will be testifying about matters that grow “naturally and directly out of research they have conducted independent of litigation, or whether they have developed their opinions expressly for purposes of testifying”; whether “the expert has unjustifiably extrapolated from an accepted premise to an unfounded conclusion”; whether “the expert has adequately accounted for obvious alternative explanations”; whether “the expert is being as careful as he would be in his regular professional work outside his paid litigation consulting”;

injury—outside the realm of litigation. It is only when the plaintiff discovers the injury that research into causation would begin, and it is also at that very point that litigation is anticipated. Indeed, every treating doctor who testifies to the plaintiff’s injuries can be said to conduct her research in anticipation of litigation.

13. See also *Wendell v. GlaxoSmithKline LLC*, 858 F.3d 1227, 1235 (9th Cir. 2017) (“[T]he district court was wrong to put so much weight on the fact that the experts’ opinions were not developed independently of litigation and had not been published. While independent research into the topic at issue is helpful to establish reliability, its absence does not mean the experts’ methods were unreliable . . . [E]xpert testimony may still be reliable and admissible without peer review and publication. That is especially true when dealing with rare diseases that do not impel published studies.”) (citations omitted).

14. See *Daubert*, 43 F.3d 1311, 1318 at n.9.

15. *Id.* at 1319.

The Admissibility of Expert Testimony

and whether “the field of expertise claimed by the expert is known to reach reliable results for the type of opinion the expert would give.”¹⁶

The *Daubert* opinion sent some mixed messages and left some unanswered questions, however. It sent mixed messages by criticizing the rigidity of the *Frye* general acceptance standard as being too “austere” and by encouraging a “flexible” inquiry, while at the same time directing trial judges to ensure that an expert’s opinion was properly derived from the “scientific method.” It clearly announced the trial judge’s obligation to find reliability by a preponderance of the evidence under Rule 104(a), but then noted that “[v]igorous cross-examination, presentation of contrary evidence, and careful instruction on the burden of proof are the traditional and appropriate means of attacking shaky but admissible evidence.”¹⁷ Of course, the trial methods referred to by the Court are the proper methods for attacking testimony—*once it has been found admissible*. But this language caused some to assume, mistakenly, that “shaky” expert opinion testimony should be admitted, with issues of reliability to be resolved by the jury as ones of weight.

Daubert also left open questions. The opinion very clearly described the standard for admitting “scientific” expert testimony, leaving open the questions of which testimony falls within this category, as well as the standard applicable to nonscientific expert opinion testimony. Finally, because the Supreme Court remanded the case to the lower courts, it did not apply the reliability standard announced in *Daubert* to a concrete set of facts, leaving the lower courts to apply the new standard in the first instance.

Some of these open questions were addressed by the Supreme Court soon after. The Court offered insight into the proper application of the *Daubert* reliability standard in *General Electric Co. v. Joiner*.¹⁸ In *Joiner*, the plaintiff proffered experts to testify that his exposure to polychlorinated biphenyls (PCBs) promoted his development of small cell lung cancer. The district court excluded the plaintiff’s experts as unreliable under *Daubert* and granted summary judgment for the defendants. On appeal, the Eleventh Circuit reversed, stating that “[b]ecause the Federal Rules of Evidence governing expert testimony display a preference for admissibility, we apply a particularly stringent standard of review to the trial judge’s exclusion of expert testimony.”¹⁹ Under this stringent standard of review, the Eleventh Circuit reversed the exclusion of the plaintiff’s experts.

On appeal, the Supreme Court first held that the abuse of discretion standard of review applies to a district court’s rulings admitting or excluding expert testimony under Rule 702 and that the Eleventh Circuit had erred in applying a

16. *In re Marriott Int’l, Inc., Customer Data Sec. Breach Litig.*, 602 F. Supp. 3d 767, 774 (D. Md. 2022).

17. 509 U.S. at 596.

18. 522 U.S. 136, 146 (1997).

19. *Joiner v. Gen. Elec. Co.*, 78 F.3d 524 (11th Cir. 1996).

more rigid standard.²⁰ Thus, appellate review of a trial court's *Daubert* ruling proceeds in two steps. The appellate court reviews de novo whether a trial court correctly applied the *Daubert*/Rule 702 framework and reviews for abuse of discretion the trial court's ruling admitting or excluding expert testimony.²¹ The Supreme Court in *Joiner* found no abuse of discretion in the trial court's exclusion of the plaintiff's experts as unreliable, emphasizing that a trial judge must ensure that an expert's *application* of her principles and methods to the facts of the case is also reliable. The experts relied on several studies, but the Court found that these studies did not support the experts' conclusions as to causation. In one study, infant mice were exposed to PCBs and contracted a different kind of cancer than that suffered by the plaintiff. Moreover, the studies could not be replicated in adult mice or in any other species. The experts' reliance on four epidemiological studies was likewise flawed. Two of the studies found no statistically significant connection between PCBs and cancer; one study did not mention PCBs; and the fourth study, which found a statistically significant connection, involved subjects who were exposed to a variety of other carcinogens. The Court concluded that there was an "analytical gap" between the experts' conclusion on causation and the studies on which they relied for their conclusion. The experts made no attempt to bridge this analytical gap by relying on any other objective or verifiable standards. The *Joiner* Court concluded that the expert testimony was merely "ipse dixit"—it is so because I am an expert and I say so. According to the Court, such testimony fails to fulfill the *Daubert* requirement that an expert's conclusions be based on an objective and verifiable methodology.

Justice Stevens dissented from this portion of the opinion, arguing that the majority had examined each study relied upon by the plaintiff's experts in a piecemeal fashion and concluded that the experts' opinions on causation were unreliable because no *one* study supported causation. Justice Stevens argued that a "weight of the evidence" approach to scientific reasoning, in which the experts relied on *all of the studies* taken together, as well as interviews of the plaintiff and an examination of his medical records to conclude that his exposure to PCBs promoted his development of lung cancer, was appropriate and acceptable under *Daubert*.²²

Daubert stated that trial courts should focus "on principles and methodology, not on the conclusions that they generate." But the *Joiner* Court clarified that a trial court should not let an opinion pass through the gate of admissibility when there is a clear disconnect between the stated methods and the conclusion reached by the expert. The *Joiner* Court explained that "conclusions and

20. 522 U.S. at 143.

21. *Kirk v. Clark Equip. Co.*, 991 F.3d 865, 872 (7th Cir. 2021).

22. *Id.* For a discussion of the "weight of the evidence" methodology, see section titled "Weight of the Evidence Analysis: Closing the Analytical Gap" below.

The Admissibility of Expert Testimony

methodology are not entirely distinct from one another” because if an expert uses a reliable method and comes to an outlier conclusion, the gatekeeper can fairly assume that there is some disconnect in the expert’s application of that methodology. According to the Court, an expert must not only be using a reliable methodology (like reviewing studies, as in *Joiner*); she must also be reliably applying that methodology to the facts at issue. As one court has stated, trial judges “may evaluate the data offered to support an expert’s bottom-line opinions to determine if that data provides adequate support to mark the expert’s testimony as reliable.”²³

Joiner did not authorize trial courts “to determine which of several competing scientific theories has the best provenance.”²⁴ Further, “*Daubert* does not require that a party who proffers expert testimony carry the burden of proving to the judge that the expert’s assessment of the situation is actually correct.”²⁵ Qualified scientific experts applying reliable methodologies might still have reasonable differences of opinion. The proponent of the evidence must show only that the expert’s conclusion has been arrived at in a scientifically sound and methodologically reliable fashion.

Another question left open by *Daubert* was whether its standards apply to expert testimony that does not purport to be scientifically based. In *Kumho Tire Co. v. Carmichael*,²⁶ the Court held that the *Daubert* reliability standard applies to nonscientific expert opinion testimony. *Kumho* involved a tire failure expert who would have testified that a tire exploded because it was defective. His mode of analysis was to investigate whether the failure might have been caused for any reason other than defect. He employed his own four-factor test to rule out other causes such as underinflation. He found that some of his self-selected factors indicated improper inflation of the tire. Still, the expert proposed to testify that improper inflation was not the cause of the tire failure. The trial judge excluded the expert’s testimony for, among other reasons, failure to satisfy the *Daubert* standards. The judge found that the expert’s methodology was subjective: it had not been peer reviewed; there was no indication of the rate of error; and there was no general acceptance of the expert’s four-factor test for determining alternative causation.

On appeal, the Supreme Court reasoned that Rule 702 requires experts to testify on the basis of “knowledge,” and that in *Daubert*, “the Court specified that it is the rule’s word ‘knowledge,’ not the modifying words like ‘scientific’, that ‘establishes a standard of evidentiary reliability.’” The Court also noted that “it would prove difficult, if not impossible, for judges to administer evidentiary

23. *Ruiz-Troche v. Pepsi Cola of P.R. Bottling Co.*, 161 F.3d 77, 81 (1st Cir. 1998).

24. *Id.*

25. See Fed. R. Evid. 702 advisory committee’s note to 2000 amendment (“When a trial court, applying this amendment, rules that an expert’s testimony is reliable, this does not necessarily mean that contradictory testimony is unreliable.”).

26. 526 U.S. 137, 147–48 (1999).

rules under which a gatekeeping obligation depended upon a distinction between ‘scientific’ knowledge and ‘technical’ or ‘other specialized’ knowledge. There is no clear line that divides the one from the others.”²⁷ The Court therefore found that the Rule 702 reliability standard applies to all expert opinion testimony—whether supported by scientific, technical, or other specialized knowledge.

The Court then examined whether a trial judge could assess nonscientific expert testimony using the specific factors that *Daubert* had listed as relevant to assessing the reliability of scientific expert testimony. The Court held that any or all of these factors might be used to evaluate nonscientific expert testimony, but found that the trial judge must have considerable leeway in deciding *how* to determine whether particular expert testimony is reliable:

Daubert’s list of specific factors neither necessarily nor exclusively applies to all experts or in every case. Rather, the law grants a district court the same broad latitude when it decides *how* to determine reliability as it enjoys in respect to its ultimate reliability determination.²⁸

The Court emphasized that the factors that a trial judge uses to determine reliability must be dependent upon the particular area of expertise that is being evaluated. The Court explained as follows:

[W]e can neither rule out, nor rule in, for all cases and for all time the applicability of the factors mentioned in *Daubert*, nor can we now do so for subsets of cases categorized by category of expert or by kind of evidence. Too much depends upon the particular circumstances of the particular case at issue.

At the same time . . . some of *Daubert*’s questions can help to evaluate the reliability even of experience-based testimony. In certain cases, it will be appropriate for the trial judge to ask, for example, how often an engineering expert’s experience-based methodology has produced erroneous results, or whether such a method is generally accepted in the relevant engineering community. Likewise, it will at times be useful to ask even of a witness whose expertise is based purely on experience, say, a perfume tester able to distinguish among 140 odors at a sniff, whether his preparation is of a kind that others in the field would recognize as acceptable.²⁹

The *Kumho* Court provided a bottom-line standard for employing the gatekeeping requirement in evaluating nonscientific expert testimony:

The objective of that requirement . . . is to make certain that an expert, whether basing testimony upon professional studies or personal experience, employs in the courtroom the same level of intellectual rigor that characterizes the practice of an expert in the relevant field.³⁰

27. *Id.* at 148.

28. *Id.* at 142.

29. 526 U.S. at 151.

30. *Id.* at 152.

The Admissibility of Expert Testimony

Turning to the tire failure expert's proffered testimony in *Kumho*, the Court found no abuse of discretion in either the trial judge's use of specific *Daubert* factors or in the ultimate decision to exclude the testimony as insufficiently reliable. The expert's assumption—that alternative causes could be ruled out by looking at a number of specific factors identified solely by the expert—was essentially subjective and unsupported by any other tire failure expert. Moreover, the expert discounted evidence that under his own test indicated that the tire had been improperly inflated. The Court concluded as follows:

Indeed, no one has argued that Carlson himself, were he still working for Michelin, would have concluded in a report to his employer that a similar tire was similarly defective on grounds identical to those upon which he rested his conclusion here. Of course, Carlson himself claimed that his method was accurate, but, as we pointed out in *Joiner*, “nothing in either *Daubert* or the Federal Rules of Evidence requires a district court to admit opinion evidence that is connected to existing data only by the ipse dixit of the expert.”³¹

Thus, the expert was properly rejected because he could not show that he used the same intellectual rigor in reaching his conclusion as would be expected of him in his professional life outside the courtroom.

Amendments to Rule 702

Rule 702 has been amended since its original enactment to clarify the trial judge's gatekeeping role announced in *Daubert*. In 2000, the rule was amended to incorporate the teachings of the Supreme Court's *Daubert* trilogy. In 2011, Rule 702 was amended as part of the restyling project that was designed to make the rules “more easily understood and to make style and terminology consistent throughout the rules.” In 2023, Rule 702 was amended again to emphasize the trial judge's obligation to find its requirements satisfied by a preponderance of the evidence and to highlight the problem of expert overstatement.³² Because the amendments to Rule 702 have codified principles governing the admissibility of expert opinion testimony found in the caselaw, pre-amendment cases may remain instructive in applying Rule 702, as amended. Judges and litigants should exercise caution in relying on pre-amendment cases, however, because the 2023 amendment was designed to correct misapplications of the rule in some federal cases.³³

31. *Id.* at 157.

32. The amendment took effect on December 1, 2023.

33. See Fed. R. Evid. 702 advisory committee's note to 2023 amendment (“... many courts have held that the critical questions of the sufficiency of an expert's basis, and the application of the expert's methodology, are questions of weight and not admissibility. These rulings are an incorrect application of Rules 702 and 104(a).”).

The 2000 Amendment to Rule 702

In 2000, Rule 702 was amended for two purposes: 1) to codify, structure, and expand upon the Supreme Court trilogy of *Daubert*, *Joiner*, and *Kumho* and its extensive progeny; and 2) to provide an extensive Committee Note to assist trial courts in exercising the gatekeeping function allocated to them under *Daubert*.³⁴ The amendment retained the preexisting standards for qualifying an expert witness and for ensuring that expert opinion testimony “will help” the jury (which were part of the original rule and were placed in a new subdivision (a)). The amendment added subsections (b)–(d) to the structure of the rule, setting forth three reliability-based standards that must be met before a challenged expert’s testimony can be admitted.

Rule 702: A Qualified Expert

Amended Rule 702 retained the requirement that an expert be qualified by “knowledge, skill, experience, training, or education.” This flexible qualification standard allows courts to qualify expert witnesses on a variety of grounds and in many disciplines. For example, in *United States v. Prothro*, a panel of the Seventh Circuit upheld the district court’s decision to admit expert opinion testimony about the make, model, and year of a kidnapper’s vehicle from surveillance videos despite the defendant’s objection that the witness lacked the requisite expertise. The appellate court found that the witness’s “position, engineering education, and nearly three decades at Ford make him abundantly qualified to opine on the appearance and identity of Ford’s products.”³⁵ In *Cedar Lodge Plantation, L.L.C. v. CSHV Fairway View I, L.L.C.*, the district court found the plaintiff’s environmental expert unqualified to offer reliable expert testimony because his experience was related to the resolution of hazardous waste matters for commercial and industrial facilities, rather than sewage systems for apartment complexes or multi-family residential communities like the one at issue in the case.³⁶ But a panel of the Fifth Circuit, in an unpublished opinion, found that the expert had “extensive experience in analysis and evaluation of environmental

34. See *In re Marriott Int’l, Inc., Customer Data Sec. Breach Litig.*, No. 19-MD-2879, 2022 WL 1323139, at *4 (D. Md. May 3, 2022) (“The advisory note to the 2000 amendments to Rule 702 (‘2000 Advisory Note’) is essential reading for judges and lawyers who undertake a *Daubert* analysis.”).

35. *United States v. Prothro*, 41 F.4th 812, 823 (7th Cir. 2022). See also *United States v. Smith*, 919 F.3d 825, 835 (4th Cir. 2019) (“twelve-year veteran of the FBI with substantial experience in drug and gang investigations (conducting controlled buys, interviewing drug dealers and gang members, using gang members as confidential sources, listening to thousands of recorded conversations, and so forth)” qualified as an expert in drug and gang terminology).

36. 753 F. App’x 191, 195–96 (5th Cir. 2018).

The Admissibility of Expert Testimony

contaminants, the area in which he was offered as an expert, including experience working on sewage systems for residential neighborhood communities.” The court held that the witness’s lack of specialization in sewage facilities for multi-family residential units did not render him unqualified.

Rule 702(a): The Expert’s Opinion “Will Help” the Trier of Fact

Rule 702 also requires that expert opinion testimony “will help” the trier of fact “to understand the evidence or to determine a fact in issue.” “[T]he ‘help’ requirement [from Rule 702] is satisfied where the expert testimony advances the trier of fact’s understanding to any degree.”³⁷ The original Advisory Committee’s note to Rule 702 quoted a law review article explaining the helpfulness requirement as follows:

There is no more certain test for determining when experts may be used than the common sense inquiry whether the untrained layman would be qualified to determine intelligently and to the best possible degree the particular issue without enlightenment from those having a specialized understanding of the subject involved in the dispute.³⁸

According to the original Advisory Committee’s note, unhelpful opinions are “superfluous and a waste of time.” In *Daubert*, the Supreme Court also characterized this requirement as one of relevance and “fit” between the opinion testimony and the dispute at issue.³⁹

United States v. King offers an example of a court evaluating an expert’s helpfulness. In *King*, the defendant in a pill mill case objected to the testimony of a medical expert regarding standards of practice in the field of pain management, guidelines for prescribing medications, and “red flags” that indicated potential painkiller abuse.⁴⁰ The Eighth Circuit affirmed the admission of the testimony even though the expert could not definitively state that any particular prescription was illegitimate. The court found that the expert’s opinion on the general operation of the clinic “advanced the trier of fact’s understanding of the clinical practices at [the clinics] and how they differed from ordinary medical facilities.” With regard to experts relying on the scientific method, it is safe to assume that

37. See, e.g., *United States ex rel. Morsell v. Symantec Corp.*, 2020 U.S. Dist. LEXIS 54847, *12 (D.D.C. 2020) (quoting 29 Charles Alan Wright & Arthur R. Miller, *Federal Practice & Procedure* § 6264.1 (2015)).

38. See Fed. R. Evid. 702 advisory committee’s note to 1973 amendment (quoting Ladd, *Expert Testimony*, 5 Vand. L. Rev. 414, 418 (1952)).

39. 509 U.S. 579, 591 (1993).

40. 898 F.3d 797, 805–06 (8th Cir. 2018).

if the testimony is relevant, it will be helpful to factfinders who are unschooled in science.

Rule 702(b): The Expert's Opinion Testimony Is Based on Sufficient Facts or Data

This subsection requires an expert to have a sufficient *foundation* for opinion testimony. It imposes a *quantitative* requirement, ensuring that an expert has enough basis supporting the opinion to which she testifies.⁴¹ If an expert has engaged in insufficient research, or has ignored obvious factors, the opinion must be excluded under this prong of the test.⁴² Put colloquially, an expert must do her homework before offering opinion testimony. A requirement that an expert have enough data to support an opinion flows naturally from the Supreme Court's demand for "reliable" opinion testimony in *Daubert*. While this foundation requirement was "well developed in the case law and in the experience of trial lawyers and judges" prior to the 2000 amendment, it was not expressly grounded in one of the federal rules.⁴³

A good example of a faulty foundation is found in *Stephens v. Union Pacific Railroad*.⁴⁴ The plaintiff in that case suffered from mesothelioma. The plaintiff's father worked for Union Pacific when the plaintiff was a young child. The plaintiff claimed that he was exposed to asbestos fibers as a child from his father's work clothes. In an effort to defeat a summary judgment motion by the defendant, the plaintiff proffered two experts who concluded that the plaintiff's exposure to asbestos fibers from his father's work clothes was a substantial factor in causing his mesothelioma. A key premise underlying both experts' opinions was that the plaintiff was exposed to significant levels of asbestos as a result of his father's work. In affirming summary judgment for the defendant, the Ninth Circuit emphasized that neither expert had any information about the degree to which the plaintiff's father was exposed to asbestos fibers in his job, or about the plaintiff's own level of exposure, if any. Without that information, the court explained that the experts had insufficient basis to conclude that the plaintiff had

41. See Fed. R. Evid. 702 advisory committee's note to 2000 amendment ("Subpart (1) of Rule 702 calls for a quantitative rather than qualitative analysis.").

42. See, e.g., *In re Incretin-Based Therapies Prod. Liab. Litig.*, No. 21-55342, 2022 WL 898595, at *1 (9th Cir. Mar. 28, 2022) ("district court properly relied on the uncontested fact that Dr. Gale did not independently review studies that had been published between 2015 and Dr. Gale's final 2019 report, all of which found no causal relationship between liraglutide use and the development of pancreatic cancer").

43. *Elcock v. Kmart Corp.*, 233 F.3d 734, 756, n.13 (3d Cir. 2000).

44. 935 F.3d 852 (9th Cir. 2019).

The Admissibility of Expert Testimony

been exposed to asbestos with any regularity. Therefore, the experts' conclusions were based on insufficient facts or data.⁴⁵

As explained by the Advisory Committee's note to the 2000 amendment, "[t]here has been some confusion over the relationship between Rules 702 and 703" and their respective roles in regulating the basis for an expert opinion. Rule 702(b) provides a *quantitative* requirement. It demands "sufficient" facts or data supporting an expert's opinion, without which the opinion would not be reliable. Rule 703 addresses the type or *quality* of the facts or data that may be relied on by an expert. It releases expert witnesses from the standard requirement of personal knowledge, allowing them to base opinions on "facts or data in the case that the expert has been made aware of or personally observed." It then addresses the situation in which an expert relies on facts or data that would be inadmissible in evidence under the Rules to form the basis of an opinion. Rule 703 provides that experts may rely on such inadmissible basis information so long as "experts in the particular field would reasonably rely on those kinds of facts or data in forming an opinion on the subject."⁴⁶ For example, in *Nicholson v. Biomet, Inc.*, a biomedical engineer offered an opinion based, in part, on medical literature and reports that constituted inadmissible hearsay.⁴⁷ Because biomedical engineers "unquestionably rely on such data from medical experts when designing medical devices compatible with the human body," the appellate court found no error in allowing the testimony.⁴⁸ Importantly, Rule 703 prohibits the proponent of the expert from disclosing inadmissible basis information to the factfinder unless the "probative value in helping the jury evaluate the opinion substantially outweighs" its prejudicial effect. In sum, Rule 702 requires a sufficient quantity of facts or data supporting an expert opinion, while Rule 703 regulates the type of data upon which an expert may rely, as well as the disclosure of inadmissible basis to the jury by the proponent.

45. *Id.*; see also *Taylor v. Bristol Myers Squibb Co.*, 93 F.4th 339, 348 (6th Cir. 2024) (expert who relies on one study and ignores contrary studies, for no justifiable reason, is relying on insufficient facts or data; court notes that the 2023 amendment to Rule 702 was intended to correct court decisions holding that the sufficiency of an expert's basis was a question of weight and not admissibility); *Wasson v. Peabody Coal Co.*, 542 F.3d 1172 (7th Cir. 2008) (expert's testimony that lessee's coal price for calculating royalties was too low based on only a single customer; as such it was not based on sufficient facts or data and was properly excluded under Rule 702(b)).

46. See, e.g., *Nicholson v. Biomet, Inc.*, 46 F.4th 757 (8th Cir. 2022) ("Biomedical engineers, such as Truman herself, unquestionably rely on such data from medical experts when designing medical devices compatible with the human body. Thus, Truman correctly used medical reports and literature as contemplated by Rule 703 to support her opinion as a biomedical engineer on the design of the M2a Magnum.").

47. *Id.*

48. *Id.*

Rule 702(c): The Testimony Is the Product of Reliable Principles and Methods

This subsection was added to Rule 702 by the 2000 amendment to reflect the heart of the *Daubert* opinion in rule text. As the Court in *Daubert* held, the trial judge must act as a gatekeeper and must find that the principles and methods utilized by an expert to formulate an opinion are reliable.

United States v. Gissantaner offers an analysis of the reliable methodology requirement.⁴⁹ The defendant in that case was charged with the unlawful possession of a firearm by a convicted felon after a witness complained that the defendant was handling a gun and a gun was found in his roommate's drawer. The handle of the gun contained DNA that emanated from multiple sources and the prosecution sought to present testimony based on STRmix, a probabilistic genotyping software program. The testimony would have shown that the software program produced a 49 million to 1 likelihood ratio for the defendant's DNA, offering "very strong support" for the conclusion that the defendant was a contributor to the DNA sample on the handle of the gun. The district court excluded the STRmix testimony under *Daubert*, finding it insufficiently reliable as a methodology.

On appeal, the Sixth Circuit conducted a careful analysis of the reliability of STRmix as a method of identifying contributing sources to multisource DNA samples, and reversed. First, the court noted that the STRmix methodology was capable of being tested and had, in fact, been tested using lab-created multisource DNA samples to evaluate the ability of the software to include and exclude sources. Next, the court explained that STRmix has been subjected to peer review because it has been the subject of over 50 peer-reviewed scientific articles. The court noted that two of three studies performed on STRmix were conducted by individuals who were not connected to the development of the software. The court examined the error rate for the STRmix program, finding that STRmix accurately excluded noncontributors to a DNA source 99.1% of the time and produced extremely low likelihood ratios for the 1% of sources that it failed to exclude. The court also noted that a national association of forensic laboratories had developed standards regarding the use of probabilistic genotyping software that were used to minimize errors associated with STRmix and other similar software programs. Finally, the court examined the general acceptance of STRmix as a method for determining contributors to a multisource DNA sample. The court found that STRmix is the "market leader" in probabilistic genotype software that is utilized by the FBI and numerous government laboratories in the U.S. and abroad. After carefully examining each of

49. 990 F.3d 457, 468 (6th Cir. 2021).

The Admissibility of Expert Testimony

the factors outlined in *Daubert*, the court found that STRmix was a reliable methodology under Rule 702(c).⁵⁰

It will often be the case that experts in a particular field employ different methods. *Daubert* and Rule 702(c) do not authorize a trial judge to choose one competing methodology over another if both have been shown to be reliable. The question is whether the methodology employed by the expert is reliable; dispute in the field is a relevant consideration, but it is not dispositive. For example, in *Ruiz-Troche v. Pepsi Cola of P.R. Bottling Co.*,⁵¹ a wrongful death action arising from a traffic accident, the defendant proffered an expert pharmacologist who concluded that the decedent had “snorted” at least 200 milligrams of cocaine thirty to sixty minutes prior to the accident. To reach this conclusion on dosage, the expert relied on a mathematical formula said to be supported by several published studies utilizing a “half-life technique.” The trial court excluded this testimony on the ground that other pharmacologists used a different technique for determining dosage. But the appellate court found an abuse of discretion. The court recognized that the scientific literature relied on by the expert “does not make an open-and-shut case” because the studies indicate that the half-life of cocaine varies among individuals due to many factors. But the court stated that the trial judge “set the bar too high” in excluding the expert’s testimony on the basis of these possible variables. The court noted that “*Daubert* neither requires nor empowers trial courts to determine which of several competing scientific theories has the best provenance. It demands only that the expert’s conclusion has been arrived at in a scientifically sound and methodologically reliable fashion.” In this case, the expert’s opinion was “premised on an accepted technique” and “embodied a methodology that has significant support in the relevant universe of scientific literature.”⁵²

Rule 702(d): The Witness Has Applied the Principles and Methods Reliably to the Facts of the Case

This subsection was added to Rule 702 in 2000 to require a trial court “to scrutinize not only the principles and methods used by the expert, but also whether

50. *Id.*; see also *United States v. Morgan*, 45 F.4th 192 (D.C. Cir. 2022) (expert’s cell tower location opinion supported by “drive testing” was based on reliable principles and methods where “drive testing technology has been relied upon, tested and reviewed for decades in the multibillion dollar wireless communications industry”).

51. 161 F.3d 77, 85 (1st Cir. 1998).

52. *Id.*; see also *Adams v. LabCorp of Am.*, 760 F.3d 1322 (11th Cir. 2014) (reversing district court’s exclusion of plaintiffs’ expert regarding the breach of a cytotechnologist’s standard of care; expert testimony was admissible notwithstanding nonblinded review where expert used a well-established classification system to identify precancerous cells).

those principles and methods have been properly applied to the facts of the case.”⁵³ This requirement was implicit in *Daubert*’s recognition of a gatekeeping role for the trial judge and explicit in *Joiner*. Although *Daubert* stated that trial courts should focus “on principles and methodology, not on the conclusions that they generate,” the *Joiner* Court explained that “conclusions and methodology are not entirely distinct from one another.” Thus, Rule 702 was amended in 2000 to make this reliable application requirement explicit in rule text. The Advisory Committee’s note to the amendment explained:

Under the amendment, as under *Daubert*, when an expert purports to apply principles and methods in accordance with professional standards, and yet reaches a conclusion that other experts in the field would not reach, the trial court may fairly suspect that the principles and methods have not been faithfully applied.

For example, in *McGill v. BP Exploration & Production, Inc.*, the plaintiff brought suit for illnesses he allegedly suffered as a result of his exposure to oil, dispersants, and decontaminants while participating in the cleanup of the *Deepwater Horizon* oil spill.⁵⁴ The district court excluded the plaintiff’s causation expert, concluding that the expert lacked critical knowledge regarding the level of the alleged contaminants harmful to humans and the level of the plaintiff’s exposure. The district court also excluded the opinion because the expert assumed that the plaintiff’s illnesses were caused by exposure because of the temporal proximity between his injuries and the exposure.

A panel of the Fifth Circuit upheld the exclusion of the plaintiff’s causation expert in an unpublished opinion. The expert relied on studies showing respiratory harm from exposure to the contaminants at issue. Still, the studies did not support the expert’s conclusion that the *plaintiff’s* illnesses resulted from his exposure because they did not show that the contaminants caused the illnesses from which the plaintiff suffered or the levels of exposure necessary to cause harm. The court explained:

[T]here is a notable analytical gap between the facts he relies on and the conclusions he reaches. Dr. Stogner’s deposition fails to address other potential causes of McGill’s illness and the method by which he rules them out. Dr. Stogner fails to analyze the conditions of exposure McGill may have experienced. . . . Dr. Stogner was unable to answer questions regarding how much time McGill spent scooping up oil, how, where, or in what quantity Corexit was used, how exposure levels would change once substances were diluted in seawater, or how McGill’s protective equipment would affect exposure.

53. Fed. R. Evid. 702 advisory committee’s note to 2000 amendment.

54. 830 F. App’x 430, 433 (5th Cir. 2020).

The Admissibility of Expert Testimony

Thus, the expert did not reliably apply his principles and methods to the facts of the plaintiff's case.⁵⁵

Importantly, as explained in the Advisory Committee's note to the 2000 amendment, the trial court's obligation to determine that an expert has reliably applied her methodology does not authorize a judge to determine the *correctness* of an expert's conclusion. As the Ninth Circuit explained in *Elosu v. Middlefork Ranch Inc.*: "Although a district court may screen an expert opinion for reliability, and may reject testimony that is wholly speculative, it may not weigh the expert's conclusions or assume a factfinding role."⁵⁶ In that case, plaintiffs seeking recovery for a fire that burned their cabin to the ground proffered a report by a fire investigator. The expert opined that the fire was caused by an open-flame pilot light that ignited combustible vapors from an oil stain that had been applied to a wooden deck the day before the fire. Although the parties stipulated to the expert's qualifications and to the reliability of his methodology, the district court excluded his causation opinion for disagreement with its "ultimate conclusions." The trial court found that the substance of the opinion was "speculative, uncertain, and contradicted by multiple eyewitness accounts." The Ninth Circuit reversed the exclusion of the expert's testimony, finding that the trial court "disregarded much of the expert's scientific analysis, weighed the evidence on record, and demanded corroboration—factfinding steps that exceed the court's gatekeeping role." The court emphasized that "Rule 702 does not license a court to engage in freeform factfinding, to select between competing versions of the evidence, or to determine the veracity of the expert's conclusions at the admissibility stage."⁵⁷

55. *Id.* In *McKiver v. Murphy-Brown, LLC*, the defendant objected to an expert's testimony that a DNA marker of hog feces could be found on the homes neighboring defendant's farm. 980 F.3d 937, 959 (4th Cir. 2020). The defendant contested the application of the DNA methodology in the case and argued that the expert's qualifications and reliance on the DNA marker could not support an opinion about odor at the neighboring property. The appellate court affirmed admission of the testimony, finding that "Pig2bac, upon which the report was based, has been used globally to demonstrate the traceability of swine fecal wastes and was applied in this case using protocols for sampling and analysis by a team of four Ph.D. holders, each with experience with field work, animal operations, and environmental health and engineering." The court further noted that the expert's DNA methodology was properly applied to support his opinion about tracing DNA compounds and that he did not offer an opinion on odor. *See also* *United States v. Morgan*, 45 F.4th 192 (D.C. Cir. 2022) (agent reliably applied "drive testing" methodology to approximate cell phone location when he confirmed that there were no changes in cell tower networks since the incident at issue before conducting the testing and subjected his draft analysis to peer review to detect and correct any application errors).

56. 26 F.4th 1017, 1020 (9th Cir. 2022). *See also* *Kennedy v. Collagen Corp.*, 161 F.3d 1226, 1230 (9th Cir. 1998) (reversing exclusion of causation expert where studies upon which expert relied offered support for his conclusion that collagen injections caused lupus and stating "[j]udges in jury trials should not exclude [expert] testimony simply because they disagree with the conclusions of the expert").

57. *Elosu*, 26 F.4th at 1020.

The 2023 Amendment to Rule 702

Rule 702 was amended in 2023 to address two concerns.⁵⁸ First, *Daubert* clearly held that Rule 104(a) governs a trial court’s admissibility decision under Rule 702. Rule 104(a) provides that the *court* must decide any preliminary question about whether evidence is admissible. In *Bourjaily v. United States*, the Supreme Court held that Rule 104(a) requires the proponent to demonstrate to the court that admissibility requirements are met by a preponderance of the evidence.⁵⁹ The Advisory Committee’s note to the 2000 amendment to Rule 702 reiterated that this preponderance standard of proof applies to the trial judge’s findings with respect to the admissibility of expert opinion testimony.⁶⁰ Despite this, many federal opinions have included incorrect characterizations of the applicable standard, stating that questions of the sufficiency of an expert’s basis and the reliability of an expert’s application are questions of weight for the jury. These cases are highly problematic because they largely abdicate the trial judge’s gatekeeping role on the important questions of the sufficiency of an expert’s basis and the reliability of an expert’s application of a methodology to the facts of the case.⁶¹

Rule 702 was amended in 2023 to correct these misstatements and to clarify the applicable preponderance standard of proof.⁶² The amendment expressly states that the proponent of expert opinion testimony must “demonstrate[] to the court” that the admissibility requirements of Rule 702 are “more likely than not” satisfied.⁶³

The Advisory Committee’s note to the 2023 amendment was careful to point out that some challenges to expert testimony will raise matters of weight rather than admissibility even under the proper Rule 104(a) standard:

For example, if the court finds it more likely than not that an expert has a sufficient basis to support an opinion, the fact that the expert has not read every single study that exists will raise a question of weight and not admissibility. But

58. The amendment became effective December 1, 2023.

59. 483 U.S. 171 (1987).

60. See Fed. R. Evid. 702 advisory committee’s note to 2000 amendment (“the proponent has the burden of establishing that the pertinent admissibility requirements are met by a preponderance of the evidence”).

61. Language incorrectly stating that the sufficiency of facts or data and reliable application are generally questions of weight and not admissibility can be found in *Loudermill v. Dow Chemical Co.*, 863 F.2d 566 (8th Cir. 1988); *Viterbo v. Dow Chemical Co.*, 826 F.2d 420 (5th Cir. 1987); and *Smith v. Ford Motor Co.*, 215 F.3d 713 (7th Cir. 2000).

62. See Fed. R. Evid. 702 advisory committee’s note to 2023 amendment (“the rule has been amended to clarify and emphasize that expert testimony may not be admitted unless the proponent demonstrates to the court that it is more likely than not that the proffered testimony meets the admissibility requirements set forth in the rule”).

63. The Advisory Committee’s note emphasizes that “there is no intent to raise any negative inference regarding the applicability of the Rule 104(a) standard of proof for other rules.” Fed. R. Evid. 702 advisory committee’s note to 2023 amendment.

The Admissibility of Expert Testimony

this does not mean, as certain courts have held, that arguments about the sufficiency of an expert's basis always go to weight and not admissibility. Rather it means that once the court has found it more likely than not that the admissibility requirement has been met, any attack by the opponent will go only to the weight of the evidence.

Second, Rule 702 was amended in 2023 to address a common application problem in which an expert relies on reliable methods and principles to generate an opinion but overstates the conclusion that can reliably be drawn from the methodology. Rule 702(d) was amended to emphasize that each expert opinion must “stay within the bounds of what can be concluded from a reliable application of the expert’s basis and methodology.”⁶⁴ The Advisory Committee note to the amendment explains that, while the problem of expert overstatement can occur in any context, it is “especially pertinent to the testimony of forensic experts in both criminal and civil cases.” For example, in *United States v. Machado-Erazo*, a cell phone location expert testified to the *precise* location of the defendant’s phone.⁶⁵ The D.C. Circuit held that it was error to permit this testimony where current cell phone location technology can locate a phone only within a *general* area.⁶⁶ In contrast, the Seventh Circuit in *United States v. Prothro* held that a forensic fiber analyst’s testimony reliably applied forensic fiber analysis methods because the analyst “acknowledged that her results could not definitively identify fibers as coming from the same source” and “disclosed that fiber analysis can ‘never’ associate ‘a single item to the exclusion of all others’ and that consistency alone ‘is not a means of positive identification.’”⁶⁷ The 2023 amendment to Rule

64. *Id.*

65. 901 F.3d 326 (D.C. Cir. 2018).

66. *Id.*; see also *United States v. Hill*, 818 F.3d 289, 295 (7th Cir. 2016) (declaring that cell site analysis expert testimony should include a “disclaimer” regarding accuracy; the expert should not “overpromise on the technique’s precision or fail to account for its flaws”; cell site testimony was admissible where the agent testified that the defendant’s phone records were “consistent” with him being at or near relevant locations at relevant times, but clarified that he could not state whether a phone was “absolutely at a specific address”).

67. 41 F.4th 812, 821 (7th Cir. 2022). The reliability of forensic fiber comparison evidence has been subject to serious criticism. See Nat’l Research Council, *Strengthening Forensic Science in the United States: A Path Forward* 162–63 (2009) (highlighting the lack of research determining the error rate of certain fiber-comparison methods). A concurring opinion in *Prothro* found that the forensic fiber analysis should have been excluded because of the government’s failure to show that it relied on reliable principles and methods under Rule 702(c). Still, the concurring opinion noted that the fiber expert attempted to refrain from overpromising on the capabilities of her analysis. *Prothro*, 41 F.4th at 835 (“To Baloga’s credit, she did not explicitly state that any two fibers ‘matched’ or claim to predict the likelihood that they came from a common source with any statistical precision. But she implied as much by repeatedly insisting during her testimony that one would not expect two random fibers to have all the common characteristics that she identified.”). For a case involving expert overstatement outside the forensic context, see *United States v. Valencia-Lopez*, 971 F.3d 891 (9th Cir. 2020). The defendant in *Valencia-Lopez* contended that he was coerced to carry drugs for a cartel. The trial court allowed a law enforcement expert to testify that the

702(d) seeks to rein in expert overstatement by directing trial judges to determine that an expert's proffered conclusions "reflect[] a reliable application of the principles and methods to the facts of the case."

Procedures for Resolving Rule 702 Motions and Objections

Trial courts enjoy significant discretion in fashioning procedures for determining the admissibility of expert opinion testimony in a given case. As the Supreme Court explained in *Kumho Tire*:

The trial court must have the same kind of latitude in deciding *how* to test an expert's reliability, and to decide whether or when special briefing or other proceedings are needed to investigate reliability, as it enjoys when it decides *whether or not* that expert's relevant testimony is reliable. Our opinion in *Joiner* makes clear that a court of appeals is to apply an abuse-of-discretion standard when it "review[s] a trial court's decision to admit or exclude expert testimony." 522 U.S. at 138–139, 118 S. Ct. 512. That standard applies as much to the trial court's decisions about how to determine reliability as to its ultimate conclusion. Otherwise, the trial judge would lack the discretionary authority needed both to avoid unnecessary "reliability" proceedings in ordinary cases where the reliability of an expert's methods is properly taken for granted, and to require appropriate proceedings in the less usual or more complex cases where cause for questioning the expert's reliability arises.⁶⁸

Accordingly, trial judges may decide *Daubert* motions at trial based on voir dire of a proffered expert witness or before trial, based on briefing and supporting documentation or after lengthy *Daubert* hearings involving testimony from proffered expert witnesses. Whether to hold a *Daubert* hearing is within the discretion of the court.⁶⁹

Although the trial court's gatekeeper review is often conducted through a *Daubert* hearing, a hearing is not required when the evidentiary record pertinent to the expert opinions is already well developed. For example, in *Miller v. Baker Implement Co.*,⁷⁰ the court found that the trial judge did not abuse discretion in excluding the plaintiff's expert testimony without a *Daubert* hearing. It noted

likelihood of a cartel using a coerced courier was "almost nil." The Ninth Circuit found that this opinion was erroneously admitted under Rule 702 because there was no showing of how the expert's experience could lead to a conclusion of such certainty.

68. 526 U.S. 137, 152 (1999) (emphasis in original).

69. See Fed. R. Evid. 702 advisory committee's note to 2000 amendment (noting that the rule "makes no attempt to set forth procedural requirements for exercising the trial court's gatekeeping function over expert testimony").

70. 439 F.3d 407, 412 (8th Cir. 2006).

that the plaintiff submitted affidavits, a detailed explanation of the proposed expert testimony, and a legal memorandum addressing the expert evidence issues. The court noted that while *Daubert* hearings “may be necessary in some cases, the basic requirement under the law is that parties have an opportunity to be heard before the district court makes its decisions.”⁷¹

Daubert Motions and Summary Judgment

In civil cases, objections to expert testimony under Rule 702 are often made in the context of a motion for summary judgment. Rule 56 of the Federal Rules of Civil Procedure requires that material cited to support or dispute a fact in the summary judgment context can be “presented in a form that would be admissible in evidence.” Therefore, a court must make admissibility determinations to resolve a motion for summary judgment. Defendants often file a *Daubert* motion in conjunction with a motion for summary judgment. They argue that plaintiff’s proffered expert opinion testimony fails to satisfy Rule 702 and that, without a necessary expert opinion to support the plaintiff on a crucial issue, the defendant is entitled to summary judgment. As the court explained in *Cortes-Irizarry v. Corporation Insular*,⁷² “[i]f proffered expert testimony fails to cross *Daubert*’s threshold for admissibility, a district court may exclude that evidence from consideration when passing upon a motion for summary judgment.” The court in that case cautioned against trying to “gauge the reliability of expert proof on a truncated record,” however. Although the trial court enjoys significant discretion in fashioning proper procedures for evaluating *Daubert* motions, courts should ensure that the proponent of the challenged expert testimony has an adequate opportunity to develop a record supporting admissibility at the summary judgment stage. A trial court may even wish to require a full presentation of the expert testimony, and rebuttal, to facilitate the admissibility determination. Courts have displayed considerable ingenuity in devising mechanisms that permit *Daubert* rulings to be made in conjunction with motions for summary judgment.⁷³

71. *Id.*; see also, e.g., *McKiver v. Murphy-Brown, LLC*, 980 F.3d 937 (4th Cir. 2020) (noting that the trial court “was entitled to rely on the parties’ materials without requiring further submissions or a *Daubert* hearing” and that the trial court’s ruling from the bench “reflected that it considered the parties’ arguments and briefing”).

72. 111 F.3d 184, 188 (1st Cir. 1997).

73. See, e.g., *Padillas v. Stork-Gamco, Inc.*, 186 F.3d 412, 418 (3d Cir. 1999) (“An *in limine* hearing will obviously not be required whenever a *Daubert* objection is raised to a proffer of expert evidence. Whether to hold one rests in the sound discretion of the district court. But when the ruling on admissibility turns on factual issues, as it does here, at least in the summary judgment context, failure to hold such a hearing may be an abuse of discretion.”); *In re Paoli R.R. Yard PCB Litig.*, 35 F.3d 717, 736, 739 (3d Cir. 1994) (discussing the use of *in limine* hearings); *Claar v.*

Credibility Determinations and Rule 702 Motions

As discussed above, trial judges must find the requirements of Rule 702 satisfied by a preponderance of the evidence before passing expert opinion testimony on to the jury. Questions sometimes arise about the judge's ability to make credibility determinations in performing this function. At first glance, it would seem impermissible for a trial judge to make a credibility assessment because it would encroach on the jury's fundamental role. But there are some special credibility determinations that may be necessary in evaluating Rule 702's reliability requirements.

The court considered the complex relationship between expert credibility and reliability in *Elcock v. Kmart Corp.*⁷⁴ The trial judge in *Elcock* held a *Daubert* hearing and determined that one of the plaintiff's experts did not pass the reliability threshold. The judge relied, in part, on the fact that the expert had engaged in criminal acts involving fraud, although this fraudulent activity was not in any way related to the expert's professional life. The Third Circuit found the trial court's reliance on the expert's prior bad acts and consideration of the expert's general credibility to be error. Still, the court made clear that a district judge is required to make some credibility choices as the Rule 702 gatekeeper:

Consider a case in which an expert witness, during a *Daubert* hearing, claims to have looked at the key data that informed his proffered methodology, while the opponent offers testimony suggesting that the expert had not in fact conducted such an examination. Under such a scenario, a district court would necessarily have to address and resolve the credibility issue raised by the conflicting testimony in order to arrive at a conclusion regarding the reliability of the methodology at issue.⁷⁵

While the court concluded that some credibility determinations would have to be made at a *Daubert* hearing, it emphasized that those determinations should be limited to testimony about how the expert reached her opinion, as opposed to witness credibility more generally. Thus, a distinction must be drawn between credibility evidence that bears directly on the expert's methods and application, and evidence on credibility that is more general.⁷⁶

Burlington N. R.R., 29 F.3d 499, 502–05 (9th Cir. 1994) (discussing the district court's technique of ordering experts to submit serial affidavits explaining the reasoning and methodology underlying their conclusions).

74. 233 F.3d 734, 750–51 (3d Cir. 2000).

75. *Id.*

76. See *Adams v. LabCorp of Am.*, 760 F.3d 1322 (11th Cir. 2014) (issues of potential bias as a result of nonblinded review and expert's philosophical bent were matters of weight for the jury); see also Edward J. Imwinkelreid, *Trial Judges—Gatekeepers or Usurpers? Can the Trial Judge Critically*

The Admissibility of Expert Testimony

If the attack on credibility has nothing to do with the expert's methods, but only with the expert's general character for truthfulness or bias, the issue of credibility should be left to the jury—the opponent can bring impeachment evidence before the jury by way of cross-examination as with any witness. As applied to the facts of *Elcock*, the credibility evidence should not have been used by the trial court because it related to acts of dishonesty and fraud completely outside the expert's work in the particular case.⁷⁷

Important Procedural Takeaways

Whether or not the trial court holds a *Daubert* hearing, it is important to keep in mind three procedural points regarding the gatekeeper function:

- The trial court must find that the proponent has or has not shown by a preponderance of the evidence that the expert's testimony satisfies the requirements of Rule 702. The trial court cannot simply refuse to decide whether the expert relied on sufficient facts and data to support her opinion, employed reliable methods and principles, and applied those methods reliably to the facts of the case at hand. Leaving these matters “to the jury” constitutes an abdication of the trial court's gatekeeper role under Rule 104(a). For example, the trial court in *McClain v. Metabolife Int'l, Inc.*⁷⁸ held a *Daubert* hearing regarding two witnesses who were providing opinions on pharmacology and chemistry. After the hearing, the trial court essentially washed its hands of the admissibility determination and sent both witnesses to the jury. The trial court explained that it did not “pretend to know enough to formulate a logical basis for a preclusionary order that would necessarily find, as a matter of law, that these witnesses cannot express to a jury the opinions they articulated to the court.”

Assess the Admissibility of Expert Testimony Without Invading the Jury's Province to Evaluate the Credibility and Weight of the Testimony, 84 Marq. L. Rev. 1, 39 (2000) (stating that an expert's prior dishonesty or misconduct should not be considered by the court when the acts are wholly unrelated to the expert's use of a particular methodology, but that a court should take such dishonesty or misconduct into account when the nexus between the acts and the expert's methodology is more direct, e.g., when the prior dishonest acts involve fraud committed in connection with the earlier phases of a research project that serves as the foundation for the expert's proffered opinion).

77. See also, *Cruz-Vazquez v. Mennonite Gen. Hosp. Inc.*, 613 F.3d 54 (1st Cir. 2010) (error to exclude expert because he was biased in favor of plaintiffs in medical cases and was generally affiliated with plaintiffs' lawyers; those considerations are for the jury in assessing the weight of the expert's testimony).

78. 401 F.3d 1233, 1239 (11th Cir. 2005).

The court of appeals found that the trial court had made an error of law by having “essentially abdicated its gatekeeping role. Although the trial court conducted a *Daubert* hearing, and both witnesses were subject to a thorough and extensive examination, the court ultimately disavowed its ability to handle the *Daubert* issues. This abdication was in itself an abuse of discretion.”⁷⁹

- The trial court must make findings on the record and set forth its reasoning on the record as well. Expressly stating that the court has or has not found the Rule 702 requirements satisfied by a preponderance of the evidence can demonstrate that the proper standard has been applied. Without such findings, there is no way for an appellate court to determine whether the gatekeeping function was properly exercised.⁸⁰
- The gatekeeping function has been exercised flexibly by some courts in the context of bench trials. While Rule 702 certainly applies in bench trials, some courts have found that the gatekeeping function can be applied more flexibly because a trial judge serving as factfinder can be trusted to give dubious expert testimony the weight it deserves.⁸¹

Recurring Issues in the Application of Rule 702

There are several recurring issues bearing on the reliability of a scientific expert’s testimony that courts have confronted in deciding motions under Rule 702. None of the issues discussed below is necessarily dispositive, but each has been considered in the cases.

79. *Id.*

80. *See, e.g.,* *Certain Underwriters at Lloyd’s, London v. Axon Pressure Prods. Inc.*, 951 F.3d 248 (5th Cir. 2020) (abuse of discretion to exclude expert testimony by a plenary order that offered no reasons for the ruling); *United States v. Yeley-Davis*, 632 F.3d 673 (10th Cir. 2011) (it was error to permit an agent to testify about how cell phone towers work without making findings on the record that the witness was qualified and his opinions were reliable); *Smith v. Jenkins*, 732 F.3d 51 (1st Cir. 2013) (remand required given the absence of any findings on the record regarding the reliability of the expert’s opinion).

81. *See, e.g.,* *Doe v. Hamilton Cnty. Bd. of Educ.*, 392 F.3d 840, 852 (6th Cir. 2004) (noting that the gatekeeper function was “designed to protect juries and is largely irrelevant in the context of a bench trial”); *United States v. Brown*, 415 F.3d 1257, 1268 (11th Cir. 2005) (Admissibility standards of Rule 702 “are more relaxed in a bench trial situation, where the judge is serving as factfinder and we are not concerned about dumping a barrage of questionable scientific evidence on a jury. . . . There is less need for the gatekeeper to keep the gate when the gatekeeper is keeping the gate only for himself.”). *But see* *UGI Sunbury v. Permanent Easement for 1.7575 Acres*, 949 F.3d 825 (3d Cir. 2020) (finding error in admitting speculative expert testimony, and expressing skepticism about the cases stating that the gatekeeping function is relaxed in bench trials).

Extrapolation and the “Analytical Gap”

In forming their opinions, some experts start from an accepted scientific fact or premise but draw a conclusion that is not necessarily supported by the accepted premise. These experts have sometimes relied on logical analysis to go from premise to conclusion. But others have simply made an unsupported leap. If the process from premise to conclusion is not itself consistent with the scientific method or other well-accepted principles, then courts tend to exclude the testimony as based on unreliable and improper extrapolation.

For example, improper extrapolation occurs when an expert concludes, without supporting research, that a substance that causes one harm also causes a different harm. In *Lust v. Merrell Dow Pharmaceuticals, Inc.*,⁸² the question was whether the plaintiff’s birth defect, hemifacial microsomia, was caused by his mother’s use of the drug Clomid. The expert based his conclusion to that effect on epidemiological studies that showed a link between Clomid and other types of birth defects. He concluded that because Clomid is capable of causing other birth defects, it also caused the plaintiff’s hemifacial microsomia. The court explained that the expert “could have convinced the district court that his methodology was nevertheless scientific if he had explained how he went about reaching his conclusions and had pointed to an objective source demonstrating that his method and premises were generally accepted by or espoused by a recognized minority of teratologists.” The court affirmed the trial court’s exclusion of the expert’s opinion, however, because he failed to do so, finding that the expert could not simply rely on epidemiological studies regarding the connection between Clomid and other birth defects to reach a conclusion about the plaintiff’s distinct birth defect. The court found that the doctor’s testimony “was influenced by litigation-driven financial incentive,” and the doctor’s premise—that a positive association between a drug and some birth defects indicates an association with other birth defects—was not recognized by even a minority of scientists.⁸³

Red flags are also raised when a scientific expert purports to testify on the basis of a methodology that she transfers from one arena to a completely different area of inquiry—at least if there is no independent research supporting the

82. 89 F.3d 594, 597 (9th Cir. 1996).

83. See also *Mitchell v. Gencorp.*, 165 F.3d 778, 782 (10th Cir. 1999) (experts could not testify that the plaintiff’s exposure to chemicals similar to benzene would affect the body in the same way as benzene; without “scientific data supporting their conclusions that chemicals similar to benzene cause the same problems as benzene, the analytical gap in the experts’ testimony is simply too wide for the opinions to establish causation”); *In re Nexium Esomeprazole*, 662 F. App’x 528, 530 (9th Cir. 2016) (opinion of orthopedic surgeon that drug Nexium could cause reduced bone mineral density and related fractures properly excluded where he “did not adequately explain how he inferred a causal relationship from epidemiological studies that did not come to such a conclusion themselves”).

transfer. Thus, in *Braun v. Lorillard Inc.*,⁸⁴ the plaintiff sought to prove that the decedent's mesothelioma was caused by smoking cigarettes with a filter made with crocidolite asbestos. The decedent's lung tissue was tested for asbestos fibers, using the standard methodologies for the testing of human tissue of "bleach digestion" and "low temperature plasma ashing." No crocidolite fibers were found. The plaintiffs then retained an expert whose specialty was testing for asbestos in *building materials* to conduct tests on the decedent's lung tissue. This expert was unaware of the methodologies ordinarily employed in testing human tissue. He used the same test that he used on building materials to test the decedent's lung tissue, known as "high temperature ashing," and found crocidolite fibers in the tissue. The expert stated that high temperature ashing was as usable on tissue as on bricks—though he had never conducted such a test on tissue before this litigation. He also admitted that the high temperature could alter the chemistry of the sample, in which case it would be impossible to tell whether asbestos fibers were crocidolite or some other kind. But he asserted that his method was far more likely to produce a false negative than a false positive.

The Seventh Circuit, in an opinion by Judge Posner, held that the expert's testimony was properly excluded as unscientific under *Daubert*. It made the following points:

A judge or jury is not equipped to evaluate scientific innovations. If, therefore, an expert proposes to depart from the generally accepted methodology of his field and embark upon a sea of scientific uncertainty, the court may appropriately insist that he ground his departure in demonstrable and scrupulous adherence to the scientist's creed of meticulous and objective inquiry. To forsake the accepted methods without even inquiring why they are the accepted methods—in this case, why specialists in testing human tissues for asbestos fibers have never used the familiar high temperature ashing method—and without even knowing what the accepted methods are, strikes us . . . as irresponsible.⁸⁵

Judge Posner provided a bottom line: *Daubert* seeks to curtail "the hiring of reputable scientists, impressively credentialed, to testify for a fee to propositions that they have not arrived at through the methods that they use when they are doing their regular professional work rather than being paid to give an opinion helpful to one side in a lawsuit."⁸⁶

All this is not to say that learning from one subject matter can never be used to form a conclusion as to a different subject matter, or that premises established in one context are completely irrelevant in another. Extrapolation by an expert is permissible when justified by scientific support or reasoning. An example of permissible extrapolation by a scientific expert arose in *Kennedy v. Collagen*

84. 84 F.3d 230, 234 (7th Cir. 1996).

85. *Id.* at 235.

86. *Id.*

*Corp.*⁸⁷ The plaintiff claimed that her use of the facial product Zyderm caused her to contract lupus, an autoimmune disorder. The plaintiff's expert rheumatologist testified to causation on the basis of peer-reviewed articles, clinical trials, product studies, an investigation conducted by the Texas Department of Health, a physical examination of the plaintiff, laboratory tests, and the temporal proximity between the plaintiff's use of Zyderm and the onset of her autoimmune disorder. None of the studies upon which the expert relied indicated that Zyderm caused lupus; but they did establish a connection between Zyderm and autoimmune disorders that were similar to lupus. Ordinarily, an analytical gap would arise if the expert were to conclude that Zyderm causes lupus simply because it causes disorders that are similar to lupus. However, the expert filled the analytical gap by relying on the finding, established in both peer-reviewed publications and clinical studies, that Zyderm induces the body to produce the *same autoimmune antibodies* that are the hallmark of lupus. The expert was not relying solely on causation of other autoimmune diseases, but also on the fact that the manner of causation was the same for lupus and for other autoimmune diseases that *were* linked to Zyderm. The court concluded that "Dr. Spindler set forth the steps he took in arriving at his conclusion in his deposition. Dr. Spindler's analogical reasoning was based on objective, verifiable evidence and scientific methodology of the kind traditionally used by rheumatologists. That is precisely what *Daubert* requires."⁸⁸

Weight of the Evidence Analysis: Closing the Analytical Gap

In *Joiner*, the Supreme Court concluded that there was an "analytical gap" between the experts' conclusion on causation and the studies upon which they relied. As discussed above, Justice Stevens dissented, arguing that the majority had examined each study relied upon by the experts in a piecemeal fashion and had concluded that the experts' opinions on causation were unreliable because no *one* study supported causation. Justice Stevens argued that a "weight of the evidence" approach to scientific reasoning, in which the experts relied upon *all of the studies* taken together, as well as interviews of the plaintiff and an examination of his medical records to conclude that his exposure to PCBs promoted his development of lung cancer, was appropriate and acceptable under *Daubert*.⁸⁹ The majority did not address this argument directly, essentially finding insufficient support for the experts' conclusion in the studies and data relied upon in that case.

87. 161 F.3d 1226, 1230 (9th Cir. 1998).

88. *Id.*

89. Gen. Elec. Co. v. Joiner, 522 U.S. 136, 152–54 (1997).

A “weight of the evidence” process of scientific reasoning has been found sufficient to support an expert’s conclusion on general causation under Rule 702 since the *Joiner* decision. Weight of the evidence methodology allows a scientific expert to reason to a conclusion even though no single report or study exists that independently supports that conclusion. As the First Circuit described it in *Milward v. Acuity Specialty Products Group, Inc.*,⁹⁰ the weight of the evidence approach to making causal connections “is a mode of logical reasoning often described as ‘inference to the best explanation’ in which the conclusion is not guaranteed by the premises.”⁹¹

The plaintiff in *Milward* brought negligence claims against chemical companies alleging that the plaintiff’s rare type of leukemia, acute promyelocytic leukemia (APL), was caused by his routine workplace exposure to benzene-containing products. The district court excluded the plaintiff’s general causation expert, finding that his proffered testimony that benzene exposure can cause APL lacked sufficient demonstrated scientific reliability. On appeal, a panel of the First Circuit reversed, finding that the plaintiff’s expert opinion on general causation satisfied Rule 702. The court noted that the plaintiff’s general causation expert relied on a “weight of the evidence” approach in concluding that benzene was capable of causing APL.

The court explained that this approach involves an inference to the best explanation that can be thought of as involving six general steps:

The scientist must (1) identify an association between an exposure and a disease, (2) consider a range of plausible explanations for the association, (3) rank the rival explanations according to their plausibility, (4) seek additional evidence to separate the more plausible from the less plausible explanations, (5) consider all of the relevant available evidence, and (6) integrate the evidence using professional judgment to come to a conclusion about the best explanation.⁹²

The court acknowledged that this method of reasoning depends upon the use of “scientific judgment” but explained: “[t]he fact that the role of judgment in the weight of the evidence approach is more readily apparent than it is in other methodologies does not mean that the approach is any less scientific.”⁹³ The court noted that the weight of the evidence approach is commonly used by scientists, and opined that “[n]o serious argument can be made that the weight of the evidence approach is inherently unreliable” under Rule 702.

The court found that the question to be resolved by a trial court is whether the approach has been properly applied in a particular case. The *Milward* court found that the district court had erred in excluding the expert’s opinion

90. 639 F.3d 11, 17–19 (1st Cir. 2011).

91. *Id.*

92. *Id.*

93. *Id.*

The Admissibility of Expert Testimony

testimony on general causation where the expert had derived his opinion “from the accumulation of multiple scientifically acceptable inferences from different bodies of evidence” and clearly employed the “same level of intellectual rigor” that he did in his academic work. The *Milward* court found that the district court erred in treating “the separate evidentiary components of Dr. Smith’s analysis atomistically”:

In Dr. Smith’s weight of the evidence approach, no body of evidence was itself treated as justifying an inference of causation. Rather, each body of evidence was treated as grounds for the subsidiary conclusion that it would, if combined with other evidence, support a causal inference. The district court erred in reasoning that because no one line of evidence supported a reliable inference of causation, an inference of causation based on the totality of the evidence was unreliable. The hallmark of the weight of the evidence approach is reasoning to the best explanation for all of the available evidence.⁹⁴

The First Circuit explained that the weight of the evidence approach may be particularly appropriate in cases like the plaintiff’s because “the rarity of APL and difficulties of data collection in the United States make it very difficult to perform an epidemiological study of the causes of APL that would yield statistically significant results.”⁹⁵ Thus, the court held that the trial court abused its discretion by concluding that the expert’s inference of causation based on the totality of the evidence was unreliable “because no one line of evidence supported a reliable inference of causation.”⁹⁶

94. *Id.* (citations omitted).

95. *Id.*

96. See also *In re Abilify* (Aripiprazole) Prod. Liab. Litig., 299 F. Supp. 3d 1291, 1311 (N.D. Fla. 2018) (“This ‘weight of the evidence’ approach to analyzing causation can be considered reliable, provided the expert considers all available evidence carefully and explains how the relative weight of the various pieces of evidence led to his conclusion.”); *Waite v. All Acquisition Corp.*, 194 F. Supp. 3d 1298, 1313 (S.D. Fla. 2016) (“This weight-of-the-evidence approach to causation requires consideration of all available scientific evidence, including epidemiology, toxicology (animal studies), cellular studies (in vitro), and molecular biology and has been validated by the courts.”). But see *In re Zolof* (Sertraline Hydrochloride) Prod. Liab. Litig., 858 F.3d 787, 796–97 (3d Cir. 2017) (“Here, we accept that the Bradford Hill and weight of the evidence analyses are generally reliable. We also assume that the ‘techniques’ used to implement the analysis (here, meta-analysis, trend analysis, and reanalysis) are themselves reliable. However, we find that Dr. Jewell did not 1) reliably apply the ‘techniques’ to the body of evidence or 2) adequately explain how this analysis supports specified Bradford Hill criteria. Because ‘any step that renders the analysis unreliable under the *Daubert* factors renders the expert’s testimony inadmissible,’ this is sufficient to show that the District Court did not abuse its discretion in excluding Dr. Jewell’s testimony.”) (citations omitted); *In re Nexium* (Esomeprazole), 662 F. App’x 528, 530 (9th Cir. 2016) (“At best, Dr. Bal analyzed three of the nine Bradford Hill factors that guide scientists in drawing causal conclusions from epidemiological studies.”). For a detailed analysis of the *Milward* decision and the weight of the evidence approach to scientific reasoning, see *Symposium: Toxic Tort Litigation: After Milward v. Acuity Products*, 3 Wake Forest J.L. & Pol’y 1 (2013).

Reliance on Anecdotal Evidence

Courts have generally found that an expert who bases a scientific opinion solely on one or even a few case studies has an insufficient basis for reliable testimony under Rule 702. There are two problems with relying on such anecdotal data, at least exclusively. First, anecdotal evidence is usually derived from insufficient sampling—a handful of select cases that do not reflect a cross section of the relevant population. Second, there is a strong possibility that anecdotal cases are not comparable to the facts of the case at bar. For example, if a number of people contract lung cancer through an alleged exposure to a toxic substance, a proper statistical study of the causative link between the toxic substance and the cancer will exclude the possibility of other sources for the injury (referred to as confounding factors), such as cigarette smoking. But a single case study, or other anecdotal evidence, is unlikely to be screened for confounding factors.

For example, in *Cavallo v. Star Enterprises*,⁹⁷ the plaintiff alleged that she suffered respiratory illness as a result of exposure to aviation jet fuel vapors that were released from the defendant's storage terminal. The plaintiff's expert toxicologist was prohibited from testifying after a *Daubert* hearing, and the court consequently granted summary judgment for the defendant. To support an opinion that the plaintiff's respiratory illness was caused by the jet fuel vapors, the toxicologist relied solely on case studies in which people who were exposed to the organic compounds in jet fuel suffered respiratory illnesses. The court held that reliance on these studies to form a conclusion was inconsistent with the scientific method. The court reasoned that "case reports are not reliable scientific evidence of causation, because they simply describe reported phenomena without comparison to the rate at which the phenomena occur in the general population or in a defined control group; do not isolate and exclude potentially alternative causes; and do not investigate or explain the mechanism of causation."⁹⁸

This is not to say that case studies are completely irrelevant to an analysis consistent with the scientific method. The problem in *Cavallo* was that the case studies were, essentially, the *only* source upon which the expert based his opinion. In contrast, scientists often use case studies to spur further research or to confirm conclusions reached in more methodical studies.⁹⁹

97. 892 F. Supp. 756, 767 (E.D. Va. 1995), *aff'd in pertinent part*, 100 F.3d 1150 (4th Cir. 1996).

98. *Id.*; see also *Norris v. Baxter Healthcare Corp.*, 397 F.3d 878, 885 (10th Cir. 2005) ("... case reports suffer from a similar failing. Case reports that state that some women with breast implants developed disease do not provide an adequate scientific basis from which to conclude that breast implants in fact cause disease. A correlation does not equal causation.").

99. See, e.g., *Wendell v. GlaxoSmithKline LLC*, 858 F.3d 1227, 1236–37 (9th Cir. 2017) ("Although case studies alone generally do not prove causation, they 'may support other proof of causation.' Here, the experts relied not just on these studies—which not only examined reported cases but also used statistical analysis to come up with risk rates—but also on their own wealth of experience and additional literature.") (citation omitted); *Cantrell v. GAF Corp.*, 999 F.2d 1007,

Reliance on Temporal Proximity

There are a number of cases in which a healthy plaintiff has become ill shortly after his exposure to a product. For example, in *Porter v. Whitehall Lab.*,¹⁰⁰ Porter came to the hospital to be treated for a fractured toe. He had no other documented health problems. He was given ibuprofen. Over the following month, the plaintiff took about thirty tablets. At the end of that month, the plaintiff was diagnosed with rapidly progressive glomerulonephritis (RPGN), a form of end-stage renal failure from which he would not recover. At the time, no studies demonstrated a link between ibuprofen use and RPGN.¹⁰¹ Porter's experts, however, testified that ibuprofen caused his RPGN. The experts essentially based their opinions on Porter's prior good health and the temporal proximity between the ingestion of ibuprofen and his renal failure. The court concluded, however, that forming a conclusion on the basis of temporal proximity alone, in the absence of some established scientific connection between substance and illness, is inconsistent with the scientific method. By relying *solely* on temporal proximity, the expert fails to consider other possible explanations that a scientist must consider before drawing a conclusion. An expert who determines causation by relying only on temporal proximity commits the classic error of confusing correlation and causation.

But this is not to say that the temporal proximity between exposure and injury can never be relied upon in reaching an opinion on causation. The result in *Porter* might well have been different if published controlled studies and/or epidemiological evidence established some connection between his ibuprofen intake and RPGN. Then the temporal proximity between exposure and injury could be used as confirmatory data. This was the case in *Kennedy v. Collagen Corp.*, above, where the expert's opinion on causation was based, in part, on the plaintiff's having contracted lupus shortly after receiving collagen injections.¹⁰² The expert could properly rely on this factor in light of the studies indicating a substantial connection between collagen and the autoimmune antibodies present

1014 (6th Cir. 1993) (expert testified that asbestos created a risk of laryngeal cancer, basing his conclusion on epidemiological evidence reported in the medical literature, and on the inordinately high incidence of persons at the plaintiffs' workplace whom the expert had personally diagnosed as having laryngeal cancer; the court held that the expert testimony was properly admitted, stating that "[n]othing in Rules 702 and 703 or in *Daubert* prohibits an expert from testifying to confirmatory data, gained through his own clinical experience, on the origin of a disease or the consequences of exposure to certain conditions").

100. 9 F.3d 607 (7th Cir. 1993).

101. There were studies suggesting a connection between ibuprofen and interstitial nephritis—a secondary kidney condition from which Mr. Porter also suffered. But there was no demonstrated connection between nephritis and RPGN. Nor did the studies showing a connection between nephritis and ibuprofen intake support causation as a result of the very modest dosage to which Porter was subjected.

102. 161 F.3d 1226, 1230 (9th Cir. 1998).

in lupus patients. As one court put it, “[a] temporal connection is entitled to greater weight when there is an established scientific connection between exposure and illness or other circumstantial evidence supporting the causal link.”¹⁰³

Insufficient Connection Between the Expert’s Opinion and the Facts in the Case: The Requirement of “Fit”

Some experts may rely on a proper methodology, but then fail to connect the methodology with the facts of the case. This is the problem of “fit.” An example of a failure to connect reliable expert testimony with the facts arose in *Harris v. Remington Arms Company, LLC*.¹⁰⁴ In *Harris*, the plaintiff brought a product liability action alleging that her Remington rifle had fired and injured her when the safety was turned off—but the trigger was not pulled. The plaintiff offered an expert witness to explain how the rifle could have fired without anyone pulling the trigger. The expert testified that the safety and the trigger mechanism bonded together after the plaintiff had engaged the safety and then stored the rifle in a cold room, causing a liquid bonding agent to solidify. According to the expert, therefore, disengaging the safety caused the trigger to be pulled simultaneously. There was no challenge to the methodology relied upon by the expert to conclude that the liquid bonding agent could solidify in cold temperatures. But this testimony did not fit the facts of the case because the plaintiff had undisputedly disengaged the safety at other times after taking the gun from the cold room, and it had not gone off. The court found that the trial court did not err in excluding the expert’s testimony, because of its “misfit with the undisputed evidence.”¹⁰⁵

103. *Curtis v. M&S Petroleum, Inc.*, 174 F.3d 661, 670 (5th Cir. 1999) (abuse of discretion to exclude the testimony of a doctor who concluded that the plaintiffs’ health problems were caused by exposure to benzene; the expert relied not only on the “strong temporal connection between the refinery workers’ exposure to benzene and the onset of their symptoms” but also on extensive scientific literature establishing a connection between benzene and the symptoms suffered by the plaintiffs). *See also* *Claussen v. M/V New Carissa*, 339 F.3d 1049 (9th Cir. 2003) (in concluding that damage to oyster beds was caused by an oil spill, the expert could rely on the temporal proximity between the spill and the damage, when combined with field studies, government reports, and exclusion of other obvious causes, to reach a reliable opinion under *Daubert*).

104. 997 F.3d 1107, 1114 (10th Cir. 2021).

105. *Id.* (“... the district court didn’t reject Mr. Powell’s methodology. The court instead concluded that Mr. Powell’s opinion testimony didn’t fit the undisputed evidence.”); *see also* *Bogosian v. Mercedes-Benz of N. Am., Inc.*, 104 F.3d 472, 479 (1st Cir. 1997). In that case, the plaintiff was run over by her car after putting it in park and exiting the vehicle. She sought to have an engineer testify about the phenomenon of “false park detent”—where the driver feels as if he or she has put the car in park, without looking at the gear shift, but the car is not actually in park. The court held that this testimony was properly excluded under *Daubert* because the testimony rested on a factual premise—that

Lack of Testing

One of the factors identified by *Daubert* for assessing the reliability of an expert's methodology pursuant to Rule 702 was whether the expert's technique or theory can be or has been tested. Courts sometimes exclude experts when they have not tested theories or techniques that are capable of being tested. For example, in *Truck Insurance Exchange v. MagneTek Inc.*,¹⁰⁶ the Tenth Circuit held that the trial court properly excluded plaintiff's expert testimony concerning the novel theory of "pyrolysis." This theory hypothesized that wood could ignite at temperatures much lower than normal under particular circumstances. The expert testimony was excluded because neither the experts nor others in the scientific community had tested the novel theory sufficiently to demonstrate its scientific reliability. The Tenth Circuit affirmed, explaining that the importance of testing as a factor in determining reliability is at its highest when an expert proposes a theory that modifies otherwise well-established knowledge about regularly occurring phenomenon—such as the temperature at which wood will ignite.¹⁰⁷

The problem of lack of testing also appears to arise frequently in product liability cases, where the plaintiff calls an expert to testify that the defendant should have designed a product differently. If the alternative design proposed by the expert has never been made and tested, either by the expert or anyone else, courts are likely to prohibit the expert from testifying that such a design was a viable and safer alternative. As the Seventh Circuit noted in *Cummins v. Lyle Indus.*,¹⁰⁸ a number of factors must go into a reliable conclusion that an alternative design should have been employed, such as the compatibility of the design with existing systems, relative efficiency, maintenance costs, ease of servicing, and the effects of price. As the *Cummins* court put it: "Many of these considerations are product-and-manufacturer-specific, and most cannot be determined reliably without testing."¹⁰⁹

the plaintiff did not look at the console shift before turning off the car—that was at odds with the plaintiff's own testimony: "The district court appropriately found it very odd that Bogosian would present an expert witness who would testify that her own unwavering testimony was incorrect."

106. 360 F.3d 1206, 1211–12 (10th Cir. 2004).

107. See also *United States v. Gabaldon*, 389 F.3d 1090, 1099 (10th Cir. 2004) (testimony of an accident reconstructionist in a murder case that the internal geometry of a car would have made it difficult for the defendant to reach the victim with enough power to inflict blows was properly excluded where he conceded that his theory could have been tested by placing defendant in the car and taking measurements, but that no such testing was done). But see *Bitler v. A.O. Smith Corp.*, 400 F.3d 1227, 1235 (10th Cir. 2004) (expert opinion as to cause of explosion was properly permitted; the fact that the expert did not test his theory about the cause of the gas leak at issue was not dispositive where his opinion was based upon "known science of copper sulfide particulate contamination as a cause of propane gas leaks").

108. 93 F.3d 362, 369 (7th Cir. 1996).

109. *Id.* at 370–71 (affirming exclusion of alternative design testimony because the alternative design had not been tested, and the expert's opinions lent themselves to testing and substantiation by the scientific method).

But the *Daubert* factors are flexible, and Rule 702 is context-specific. There is no absolute requirement that a conclusion or theory be tested to be admissible. In the specific area of design safety, the feasibility of testing and the expense of litigation must be taken into account. For example, a design engineer may seek to testify that a car could have been designed more safely. A court should not exclude the testimony solely because the expert has not built a car employing the alternative design. A good example of the proper approach to testing after *Daubert* is found in *Tassin v. Sears, Roebuck & Co.*¹¹⁰ Tassin was injured while operating a power saw and proffered an engineer who concluded that alternative designs were safer and that the defendant failed to provide adequate warnings. The district court analyzed the admissibility of this testimony as follows:

It may well be that an engineer is able to demonstrate the reliability of an alternative design without conducting scientific tests, for example, if he can point to another type of investigation or analysis that substantiates his conclusions. For example, an expert might rely upon a review of experimental, statistical, or other technical industry data, or on relevant safety studies, products, surveys, or applicable industry standards. He could also combine any one or more of these methods with his own evaluation and inspection of the product based on experience and training in working with the type of product at issue. The expert's opinion must, however, rest on more than speculation, he must use the types of information, analyses and methods relied on by experts in his field, and the information that he gathers and the methodology that he uses must reasonably support his conclusions. If the expert's opinions are based on facts, a reasonable investigation, and traditional technical/mechanical expertise, and he provides a reasonable link between the information and procedures he uses and the conclusions he reaches, then [the opinion may be admitted despite a lack of testing].¹¹¹

Applying these standards to the facts, the *Tassin* court excluded the expert's testimony on certain alternative designs based on parts that he had never tested or even seen, and the safety of which was not supported by the tests of others or by any relevant literature. However, the court held testimony as to another alternative design admissible, where the expert had actually conducted some testing and where the safety of the product received support from the relevant literature. After finding sufficient reliable support to admit this opinion under Rule 702, the court explained that the expert's failure to test more systematically and extensively presented a question of the weight to be afforded to the opinion. Finally, the court held the expert's conclusion as to inadequate warnings to be admissible. The alternative warnings suggested by the expert had not been scientifically tested. But the court found that testing as to warnings (as opposed to testing alternative designs) was not critical where the expert had

110. 946 F. Supp. 1241, 1247–48 (M.D. La. 1996).

111. *Id.*

substantial experience in both product design and in preparing product manuals and warnings.

The Problem of Overstatement

Some expert witnesses may utilize a methodology that is sufficiently reliable and that may be applied helpfully to the facts of a particular case but may err in their application by overstating the findings that the methodology is capable of producing (particularly during trial testimony). Overstatement is especially endemic in the testimony of forensic experts. Over the past two decades, various scientific bodies have raised serious questions about the reliability of forensic techniques that had theretofore been admitted without much controversy. The two most important reports are:

1. National Research Council of the National Academy of Sciences (NAS), *Strengthening Forensic Science in the United States: A Path Forward* (2009). This detailed report essentially concludes that with one exception, none of the forensic techniques currently employed—such as ballistics, handwriting, and hair identification—meet the standards of science. The exception is identification of DNA from a single source.
2. The President’s Council of Advisors on Science and Technology (PCAST), *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods* (2016). This report provides an exhaustive analysis of why forensic feature comparison methods (specifically ballistics, handwriting, fingerprinting, footwear, hair, and DNA analysis from complex mixtures) have not been validated, and how at least some of them can be strengthened so that they have validity. Particular attention is given to the problem of experts overstating their results—e.g., making statements that fingerprint identification has a “zero rate of error” when in fact the process is based on subjective judgment that by definition carries a rate of error.

Despite these important reports, most courts have admitted forensic expert opinions as reliable technical testimony under *Kumho* even though they are not based on the scientific method.¹¹² Some courts, though, have rightly focused on

112. See, e.g., *United States v. Baines*, 573 F.3d 979, 989 (10th Cir. 2009) (conceding that there are “multiple questions regarding whether fingerprint analysis can be considered truly scientific in an intellectual, abstract sense” but concluding that “nothing in the controlling legal authority we are bound to apply demands such an extremely high degree of intellectual purity” and that “fingerprint analysis is best described as an area of technical rather than scientific knowledge”); *United States v. Johnson*, 2019 U.S. Dist. LEXIS 39590 (S.D.N.Y.) (finding ballistics identification to be admissible, and claiming that the NAS and PCAST reports impose “demanding scientific standards” that “require a level of certainty and infallibility not properly applied in a courtroom”).

preventing forensic experts from *overstating* their results—for example, prohibiting such experts from testifying that they are “certain” that there is a “match,” or that the rate of error is “zero” or “infinitesimal.” The leading case on preventing overstatement is *United States v. Glynn*,¹¹³ in which the district court found that because ballistics identification is based on subjective judgment, the expert could not testify that he was a “scientist” or that the rate of error for his identification was zero. The court allowed the expert to testify but held that he could state only that it was “more likely than not” that the bullet fragment came from the defendant’s gun—as that was all the certainty that the subjective methodology could support. The trial judge explained that it is particularly important for judges to regulate expert overstatement as part of the gatekeeping role because:

[O]nce expert testimony is admitted into evidence, juries are required to evaluate the expert’s testimony and decide what weight to accord it, but are necessarily handicapped in doing so by their own lack of expertise. There is therefore a special need in such circumstances for the Court, if it admits such testimony at all, to limit the degree of confidence which the expert is reasonably permitted to espouse.¹¹⁴

While some other courts have followed suit and have sought to control overstatement in expert testimony,¹¹⁵ many others have not.¹¹⁶ Because the problem of overstatement has not been well regulated by the courts, the Advisory Committee on Evidence Rules, starting in 2017, began considering whether Rule 702

113. 578 F. Supp. 2d 567, 575 (S.D.N.Y. 2008).

114. *Id.*

115. *See, e.g., United States v. Machado-Erazo*, 901 F.3d 326 (D.C. Cir. 2018) (error for a cell phone location expert to testify to the precise location of the defendant’s phone, where the current technology can locate a phone only within a general area); *United States v. Parker*, 871 F.3d 590 (8th Cir. 2017) (trial court prohibited the expert from testifying that she was “100% sure” or “certain” that the relevant guns matched the relevant shell casings); *United States v. Hill*, 818 F.3d 289, 295 (7th Cir. 2016) (declaring that cell site analysis expert testimony should include a “disclaimer” regarding accuracy; the expert should not “overpromise on the technique’s precision or fail to account for its flaws”; cell site testimony was admissible where the agent testified that the defendant’s phone records were “consistent” with him being at or near relevant locations at relevant times, but clarified that he could not state whether a phone was “absolutely at a specific address”); *United States v. White*, 2018 U.S. Dist. LEXIS 163258 (S.D.N.Y.) (finding proposed ballistics testimony was admissible, subject to the limitation that the expert could not testify to any specific degree of certainty that there was a ballistics match between the firearms seized from the defendant and those used in the various shooting incidents).

116. *See, e.g., United States v. Straker*, 800 F.3d 570 (D.C. Cir. 2015) (fingerprint expert allowed to testify that there was a “zero rate of error in the methodology”); *United States v. Hunt*, 2020 WL 2842844 (W.D. Okla. 2020) (the defendant asked the court “to place limitations on the Government’s firearm toolmark experts because the jury will be unduly swayed by the experts if not made aware of the limitations on their methodology”; the court allowed the expert to testify to a “reasonable degree of certainty” even though the Department of Justice standards do not permit its experts to testify in those terms).

The Admissibility of Expert Testimony

should be amended to prohibit overstatement by experts. After careful consideration, the committee concluded that overstatement was already within the court's gatekeeping responsibility under Rule 702, as amended in 2000, because an expert who overstates a conclusion is not reliably applying her basis and methodology. But the committee determined that the obligation to regulate overstatement should be highlighted for the courts with a modest amendment to Rule 702(d). The 2023 amendment to Rule 702(d) discussed above requires the court to find that "the expert's opinion reflects a reliable application of the principles and methods to the facts of the case."¹¹⁷ The amendment instructs trial courts to carefully evaluate the actual opinion rendered—the opinion itself and not just the process of applying a methodology—to make sure that it does not go beyond what the expert can reliably conclude. The Committee Note to the 2023 amendment emphasizes that the focus of the amendment is on preventing experts from overstating their opinions:

Rule 702(d) has . . . been amended to emphasize that each expert opinion must stay within the bounds of what can be concluded from a reliable application of the expert's basis and methodology. Judicial gatekeeping is essential because just as jurors may be unable, due to lack of specialized knowledge, to evaluate meaningfully the reliability of scientific and other methods underlying expert opinion, jurors may also lack the specialized knowledge to determine whether the conclusions of an expert go beyond what the expert's basis and methodology may reliably support.

While the Committee Note voices objection to testimony about a "zero rate of error," the intent of the amendment is to require court scrutiny of any attempt by a forensic expert to testify to greater certainty than is supported by the methodology. A common and problematic form of expert overstatement occurs when an expert witness purports to testify "to a reasonable degree of scientific certainty." Expert overstatement was a significant focus of the PCAST report on forensic feature-comparison methods.¹¹⁸ In particular, the PCAST report recommended that courts "should never permit scientifically indefensible claims, such as . . . proof to a reasonable degree of scientific certainty."¹¹⁹ A report from the National Commission on Forensic Sciences similarly decried this type of expert overstatement and proposed that courts should forbid experts from stating their conclusions to a "reasonable degree of [field of expertise] certainty," because that term has no scientific foundation. The Department of Justice (DOJ) has adopted testimonial protocols that prohibit the use of the "reasonable degree of certainty" language for certain feature-comparison experts, and that impose important limitations on testimony regarding rates of error. For example, a DOJ

117. The amendment took effect on December 1, 2023.

118. President's Council of Advisors on Science and Technology, *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods*. (Sept. 20, 2016).

119. *Id.* at 145.

directive regarding toolmark testimony provides, in pertinent part: “An examiner shall not use the expressions ‘reasonable degree of scientific certainty,’ ‘reasonable scientific certainty,’ or similar assertions of reasonable certainty in either reports or testimony, unless required to do so by a judge or applicable law.”¹²⁰ It is for this reason that the Advisory Committee Note to the 2023 amendment to Rule 702 specifically explains: “Forensic experts should avoid assertions of absolute or one hundred percent certainty—or to a reasonable degree of scientific certainty—if the methodology is subjective and thus potentially subject to error.” Courts should thus prohibit, where possible, the use of problematic testimony to a “reasonable degree of scientific certainty” notwithstanding significant past federal precedent in which such testimony was permitted or even required.¹²¹

Experts Relying on Technical or Other Specialized Knowledge

As discussed above, the Supreme Court’s *Kumho* decision clarified that Rule 702 and the *Daubert* reliability standard apply to expert opinion testimony based upon technical or other specialized knowledge, as well as to scientific testimony. Even though the gatekeeper function applies to nonscientific expert testimony, it stands to reason (as the Court in *Kumho* recognized) that the specific *Daubert* factors will often have to be rejected or modified when reviewing most nonscientific expert testimony. A court should not exclude an expert on automobile repair on the ground that he has not published his research in a peer-reviewed journal. A court should not reject the testimony of a sociologist simply because she cannot be definitive about a possible rate of error in her findings.

120. United States Department of Justice Uniform Language for Testimony and Reports for the Forensic Firearms/Toolmarks Discipline Fracture Examination 3–4, <https://perma.cc/P6D4-BCC9>.

121. There can be a complication in rejecting the reasonable degree of certainty standard in civil cases, however. Some states appear to require a reasonable certainty standard as a matter of state substantive law—which is controlling in diversity cases, assuming that in fact it is substantive. See, e.g., *Miranda v. County of Lake*, 900 F.3d 335 (7th Cir. 2018) (“In Illinois, proximate cause must be established by expert testimony to a reasonable degree of medical certainty.”); *Day v. United States*, 865 F.3d 1082 (8th Cir. 2017) (under Arkansas law, a medical expert must testify that “the damages would not have occurred” without the defendant’s negligence; expert’s opinion “must be stated within a reasonable degree of medical certainty or probability”). For this reason, the advisory committee’s note to the 2023 amendment to Rule 702 explains that the “amendment does not, however, bar testimony that comports with substantive law requiring opinions to a particular degree of certainty.”

The Admissibility of Expert Testimony

An example of this necessary flexibility is found in *Tyus v. Urban Search Management*.¹²² The plaintiffs in *Tyus* alleged that advertising for a rental building violated the Fair Housing Act because it targeted only white lessees. The plaintiffs proffered a social science expert who would have testified to how an all-white advertising campaign affects African Americans. The court declared that the central teaching of *Daubert*—that expert testimony “must be tested to be sure that the person possesses genuine expertise in a field and that her court testimony adheres to the same standards of intellectual rigor that are demanded in her professional work”—was fully applicable to the testimony of experts in the social sciences. However, recognizing the need for a flexible application of *Daubert*, the court noted the following caveat:

It is true, of course, that the measure of intellectual rigor will vary by the field of expertise and the way of demonstrating expertise will also vary. Furthermore, we agree . . . that genuine expertise may be based on experience or training. In all cases, however, the district court must ensure that it is dealing with an expert, not just a hired gun.

The *Tyus* court found that the trial court erred in excluding the expert, because his research was based on peer-reviewed articles, and his “focus group” method was a well-accepted methodology in the field of social science. In other words, the expert brought the same intellectual rigor to his courtroom testimony as he employed in his life as a social scientist.

Where expert testimony is not adaptable to being put through the wringer of falsifiability, publication, and peer review, the trial judge still has the obligation to determine whether the testimony is properly grounded, well reasoned, and not speculative. If there is a well-accepted body of learning, experience, and basic assumptions in the field, then the expert’s testimony should be grounded in that learning and experience to be reliable. The more subjective and controversial the expert’s inquiry, the more likely the testimony is to be excluded as unreliable. The opinion in *Frymire-Brinati v. KPMG Peat Marwick*¹²³ is instructive. In this action for securities fraud, the plaintiff’s expert, an accountant, was allowed to testify that a Peat Marwick audit had improperly certified that property interests had a certain value when in fact they were worth much less. To reach this conclusion, the accountant used a discounted cash flow analysis, by which he assessed value solely on the basis of net, rather than potential, cash flow. He did not explain why he failed to consider potential cash flow, even though standard principles of valuation find it to be an important factor. Reversing a judgment for the plaintiff, the court held that the trial court abused its discretion in admitting the expert’s valuation, because the expert’s

122. 102 F.3d 256, 263 (7th Cir. 1996).

123. 2 F.3d 183, 188 (7th Cir. 1993).

methodology was faulty: by failing to consider potential cash flow, the expert's methodology would lead to the conclusion that "raw land is worthless and that a large office building in the final stages of construction also has no value even though it is fully leased out and could be sold for a hundred million dollars." The Seventh Circuit held that the expert's testimony lacked validity under *Daubert*, where his litigation methodology would not have been acceptable for clients outside the courtroom.

A particular problem after *Kumho* involves the nonscientific expert who testifies solely on the basis of experience. The example from *Kumho* is the perfume tester who sniffs a substance and identifies its brand and provenance. When asked how he can make such a specific identification, he responds, "I know it when I smell it and I smell it a lot." Should he be permitted to testify solely on this conclusory experience-based analysis? Put another way, must the trial court take the expert's word for it? The Court in *Kumho* implies that it is inconsistent with the gatekeeper function to allow an experience-based expert to testify without any explication of *how* he or she comes to a conclusion and *why* he or she believes it is correct. This implication—that experience-based experts must justify their conclusions as reliable—was codified in the 2000 amendment to Rule 702.

For example, in *United States v. Rodriguez*,¹²⁴ law enforcement officers testified as experts interpreting intercepted phone calls. But the officers testified about uncommon terms or phrases that they encountered for the first time in the investigation. They relied on experience, but the court found error in admitting their testimony, because "the officers generally did not offer an explanation for how they arrived at their interpretations, nor did the court require them to provide one." Any explanations the officers did offer for arriving at their interpretations "failed to evince indicia of reliability or methodological rigor." The court noted that "an officer's qualifications, including his experience with investigations and intercepted communications, are relevant but not alone sufficient to satisfy Federal Rule of Evidence 702" because "Rule 702 requires district courts to assure that an expert's methods for interpreting the new terminology are both reliable and adequately explained."¹²⁵ To be sure, law enforcement officers are frequently permitted to give expert testimony regarding the meaning of code words, particularly in gang- and drug-related cases, so long as they offer the requisite explanation of their methodology.¹²⁶

124. 971 F.3d 1005, 1018 (9th Cir. 2020).

125. *Id.*; see also *United States v. Hermanek*, 289 F.3d 1076 (9th Cir. 2007) (expert testimony translating purported coded conversation should have been excluded; although the government made an adequate showing of the witness's experience, the witness did not explain in detail the methods he used to arrive at his interpretation of words that he was not familiar with before the investigation).

126. See *United States v. Wilson*, 484 F.3d 267, 275 (4th Cir. 2007) (explaining that "courts of appeals have routinely held that law enforcement officers with extensive drug experience are

Conclusion

Exercising the gatekeeping function under *Daubert* and Rule 702 is sometimes daunting and often requires a substantial expenditure of judicial resources. But the alternative—leaving questions of expert reliability to the jury—is inconsistent with Rule 702 and established Supreme Court precedent. Thus, the trend over time has been for judges to exercise closer scrutiny of expert testimony, and this trend may be expected to continue as evidence in litigation becomes more complex.

qualified to give expert testimony on the meaning of drug-related code words” and affirming admission of such testimony where expert carefully “summarized his methodology”).

How Science Works

MICHAEL WEISBERG AND ANASTASIA THANUKOS

Michael Weisberg, Ph.D., is Bess W. Heyman President's Distinguished Professor of Philosophy and Deputy Director of Perry World House, University of Pennsylvania.

Anastasia Thanukos, Ph.D., is principal editor of the University of California Museum of Paleontology, University of California, Berkeley, and editor of the Understanding Science project.

CONTENTS

Introduction, 49

- What Is Science: “Textbook” Science vs. the Practice of Science, 50
- An Example of Science in Practice: Connecting CFCs to Ozone Depletion, 52
- The Science Flowchart, 54
- The Science Flowchart and Admissibility, 56

Key Traits of Science, 60

- Science Investigates the Natural World and Natural Explanations, 61
- Science Investigates Testable Hypotheses, 62
- Science Responds to Evidence, 63
- Science Does Not *Prove* Hypotheses, 63
- Science Is Carried Out by a Community that Holds Members to Norms, 65

Nonscience, Pseudoscience, Bad Science, and Wrong Science, 67

Science as a Human and Community Endeavor, 69

- Peer Review, 72
- Bias, 75
- Misconduct, 77
- Correction and Retraction, 78
- Scientific Expertise, 80

Testing Hypotheses, 82

- Scientific Methodologies and Data, 82
 - Experimental and Nonexperimental Methods, 83
 - Qualitative Data, Quantitative Data, and Mixed Methods, 87
 - Modeling, 90
- Correlation and Causation, 92
- Assumptions and Auxiliary Hypotheses, 93
- Replication and the “Replication Crisis”, 94

Achieving Scientific Consensus, 96

- Systematic Reviews and Meta-analyses, 99
- Not Necessarily Consensus, 100

Myths About Science, 102

Reference Manual on Scientific Evidence

Science and the Law, 106

Assessment of Evidence, 107

Precedent and Self-Correction, 108

Conclusion, 109

Glossary of Terms, 110

FIGURES

1. Science is complex, iterative, dynamic, and social. It does not proceed according to a linear, step-by-step process, 55
2. Relationships among the *Daubert* factors and the process of science, 58
3. Indicators of scientific consensus. Scientific consensus is formed based on a body of evidence, not a single study or investigation, 97

TABLE

1. Textbook Science vs. Science in Practice, 51

Introduction

Science typically appears in the courtroom as scientific evidence presented by witnesses deemed experts by the courts. Such evidence may be considered by a jury or judge, depending on the type of case. Judges are the gatekeepers for this evidence, determining what information and testimony will be allowed to enter the court and influence the outcome of a trial, a challenging and essential role for several reasons. The broad public, from which juries are selected, has a limited understanding of many basic, well-established, and uncontroversial scientific concepts.¹ Yet in a trial, a jury may have to weigh evidence regarding advanced scientific knowledge that is still emerging and about which lines of evidence may conflict or be inconclusive. Complicating matters, public relations campaigns have misled the public about the true state of scientific consensus regarding certain scientific issues.² Further, scientific expertise and the trappings of science can be made to seem persuasive, even when there is little substance to them, making it more difficult for a jury to come to a just decision when presented with shaky evidence about which they have limited scientific understanding. Thus, the responsibility to ensure that the testimony the jury hears is scientifically sound and meets minimal standards of reliability is an important one.

The purpose of this reference guide is to provide a nuts-and-bolts look at how science works today to frame subsequent reference guides of this reference manual and to inform judges' consideration of proffered scientific evidence and their decisions about what testimony to allow in a trial. Other reference guides will help judges assess questions of law (e.g., What opinions have shaped interpretation of the Federal Rules of Evidence?; see *The Admissibility of Expert Testimony*, in this manual) and questions specific to scientific disciplines (e.g., Does bitemark evidence emerge from "good" science and should it be permitted at trial?; see *Reference Guide on Forensic Feature Comparison Evidence*, in this manual). In this reference guide, we provide background and guidance to inform questions that turn on the processes and nature of science writ large: Does peer review ensure that a study is scientifically sound? What qualifies someone as an expert within the scientific community? Should the sponsor of scientific research (e.g., a government agency versus a company) factor into interpretation of its findings? Fairly evaluating such questions requires an understanding of the nature and process of science that goes beyond media and textbook portrayals. The view of science presented below is based on current research and thinking in an academic discipline that studies the theory and processes of scientific knowledge building: the philosophy of

1. Pew Research Center, What Americans Know About Science (March 2019).

2. Naomi Oreskes & Erik M. Conway, *Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Global Warming* (2010) [hereinafter Oreskes & Conway 2010].

science.³ We focus in this reference guide on the Western conception of science as it currently operates, recognizing that science is a social and cultural practice that has changed and will continue to change over time.

What Is Science: “Textbook” Science vs. the Practice of Science

Science is both a body of knowledge and the process for building that knowledge based on evidence acquired through observation, experiment, and simulation. The term is accurately applied to knowledge on a wide variety of topics and to diverse lines of inquiry. These can include familiar disciplines in the natural sciences, like medicine and physics, as well as the social and behavioral sciences. Issues involving economics, history, education, politics, social trends,

Clarifying vocabulary: Scientific knowledge

Misconceptions about and alternative definitions of the terms *hypothesis*, *theory*, *model*, and *law* abound. Herein, we use the following definitions for these elements of scientific knowledge:

Hypothesis—a potential explanation for a natural phenomenon, regardless of the extent to which the explanation has been investigated or the amount of evidence supporting or refuting it

Theory—a concise, coherent, and predictive explanation for a broad range of natural phenomena that integrates and makes sense of many hypotheses

Model—a mathematical representation of hypothesized mechanisms and interactions within a system that can be used to investigate the impact of parameter changes on predicted system outcomes

Law—an often-mathematical statement of the relationship among observable phenomena

Note that within and outside of science, these terms are used inconsistently, and their popularity has risen and fallen in different disciplines and at different points in history. For further discussion of these terms, see sections titled “Not Necessarily Consensus” and “Myths About Science” below.

3. Readers interested in an accessible introduction to this discipline, its history, and current perspective on the nature and process of science are referred to Peter Godfrey-Smith, *Theory and Reality: An Introduction to the Philosophy of Science* (2003) [hereinafter Godfrey-Smith 2003].

human opinions, and much more can be investigated with scientific rigor. This broad view of the sorts of topics within the purview of science aligns with the wide variety of topics addressed by later reference guides in this manual. In short, what makes a unit or body of knowledge scientific is the process by which it was established, not the topic it concerns.

What is this special process that makes an inquiry or investigation scientific? How do we generate scientific knowledge? Textbooks and media often present a sterile and simplified, but commonly held, view of the process of science, sometimes called the scientific method,⁴ which goes something like this: First, a scientist forms a hypothesis, a potential explanation for a natural phenomenon. For instance, a scientist might hypothesize that a particular class of chemicals is causing a hole in the ozone layer. The scientist performs an experiment to test the idea, perhaps by releasing the chemical in an ozone-filled chamber and measuring changes in the gases present in the chamber. The scientist reaches a conclusion about the hypothesis—that it is correct because ozone decreased significantly in the presence of the chemical. The results of the experiment are then written up and published. Other scientists read the article and mainly agree with the conclusions, and, voila, a new scientific fact is established. This simplified view correctly calls out an essential feature of scientific knowledge building—that it is based on evidence from the natural world that can stand up to outside scrutiny—but is misleading in other ways.

Table 1. Textbook Science vs. Science in Practice

Common representation of the scientific method	Characteristics of science in practice
Scientific investigations follow a codified step-by-step method.	Scientists engage in many different activities in different orders as they pursue evidence and explanations.
Scientists work in isolation.	Most modern science is collaborative and depends on social interactions within the scientific community.
Scientists use experiments to gather evidence.	Scientists use different methods to gather evidence; an experiment is just one of these methods.
Scientific studies reach a firm conclusion.	Scientific conclusions are always revisable if warranted by the evidence.
A single convincing study leads to scientific consensus on a hypothesis.	Scientific consensus is generally reached over the course of multiple studies conducted by different groups pursuing different lines of inquiry.

4. William F. McComas, *The Principal Elements of the Nature of Science: Dispelling the Myths, in The Nature of Science in Science Education: Rationales and Strategies* 53 (1998) [hereinafter McComas 1998].

An Example of Science in Practice: Connecting CFCs to Ozone Depletion

Although scientific inquiry always involves a reliance on evidence and has certain other traits (see section titled “Key Traits of Science” below), there is no one scientific method.⁵ For example, while scientists have reached consensus on the hypothesis that human production of a certain class of chemicals, chlorofluorocarbons (CFCs), depletes the ozone layer, the actual story is not a simple one. We shall tell this story here, and then refer to it, as well as to other scientific studies that have made their way into the courtroom, throughout this reference guide to illustrate key points about the modern practice of science.

In the 1970s, a medical researcher, curious about the source of the haze near his home, started trying to detect chemicals in the air that come exclusively from human activities—in this case, CFCs. He found them everywhere, even in Antarctica, far from where they were being produced.⁶ The findings were presented at a conference, where they caught the attention of another scientist, Sherwood Rowland, who shared it with his postdoctoral researcher, Mario Molina. Molina then turned to the published literature to figure out what might happen to these molecules once released into the atmosphere.

Molina and Rowland published an article putting forth the hypothesis that CFCs deplete our protective ozone layer, but their evidence did not come from an experiment or even any new observations.⁷ Instead, the pair brought together measurements collected by other researchers, calculations, and established knowledge about basic and atmospheric chemistry, arguing that if all these other ideas and estimates were true, a logical outcome is that CFCs pose a threat to the ozone layer. Later, experiments were performed that suggested the chemical reactions that Rowland and Molina reasoned *should* happen actually *did* happen. Other scientists incorporated these ideas into their models of the atmosphere and made predictions about what should be observed if Rowland and Molina’s hypothesis were correct. Still other groups collected atmospheric evidence to test the predictions made by the models. Meanwhile,

5. Philip Kitcher, *The Advancement of Science: Science Without Legend, Objectivity Without Illusions* (1995) [hereinafter Kitcher 1995]; William C. Wimsatt, *Re-engineering Philosophy for Limited Beings: Piecewise Approximations to Reality* (2007) [hereinafter Wimsatt 2007]. For an argument that *Daubert v. Merrell Dow Pharmaceuticals, Inc.* is partly founded on the myth of the scientific method, see Sheila Jasanoff, *Law’s Knowledge: Science for Justice in Legal Settings*, 95 Am. J. of Pub. Health s49–s58 (2005), <https://doi.org/10.2105/AJPH.2004.045732> [hereinafter Jasanoff 2005].

6. J. E. Lovelock, R. J. Maggs, & R. J. Wade, *Halogenated Hydrocarbons in and Over the Atlantic*, 241 Nature 194–96 (1973), <https://doi.org/10.1038/241194a0> [hereinafter Lovelock et al. 1973].

7. Mario J. Molina & F. Sherwood Rowland, *Stratospheric Sink for Chlorofluoromethanes: Chlorine Atom-Catalyzed Destruction of Ozone*, 249 Nature 810–12 (1974), <https://doi.org/10.1038/249810a0> [hereinafter Molina & Rowland 1974].

the CFC industry backed another scientist to oppose the hypothesis, and Rowland and Molina checked some of the old measurements on which they had based their hypothesis and found them to be inaccurate. After correcting those numbers, models were updated, compared, and updated again. More data were collected and eventually, eight years after the hypothesis was first published, researchers discovered a thinning of the ozone layer over Antarctica much more extreme than expected.⁸ This led to more revisions of the hypothesis, a few additional twists and turns, and eventually a Nobel Prize, the Montreal Protocol, which phased out CFC production, and today, a recovering ozone layer. Concurrently, ozone depletion has made its way into the courts as established science. For example, the National Resource Defense Council (NRDC) brought suit against the Environmental Protection Agency (EPA) for making decisions that violated the Montreal Protocol. The NRDC was found to have standing because of the increased risk of skin cancer that NRDC members would experience as a result of ozone destruction.⁹ Ozone depletion was treated as a fact in that case, reflecting the scientific consensus on the issue.

In this case, science worked exactly as it ought. Hypotheses were sifted through, scrutinized by different parties with different interests, compared against multiple lines of evidence, and iteratively modified, ultimately leading to reliable and actionable scientific knowledge. Such examples are typical of science as it is actually practiced: Rarely do scientists apply a simple linear recipe, rely on a single critical experiment, or reach an immediate and firm conclusion. While the complex and iterative processes that went into establishing depletion of the ozone layer by CFCs are commonplace in science, the speed with which societal and political action followed scientific consensus in this case may be unusual.¹⁰ We also note that this example of scientific consensus building played out over many years, and that, for good reason, much of the scientific testimony presented at trial will concern hypotheses with a much lower degree of certainty and that the scientific community has had less time to scrutinize. Nevertheless, throughout this reference guide the CFC/ozone example will serve as a valuable reminder of how science generally works, as well as the power of science to spur real-world change—a function that is often mediated by the courts and so further highlights the responsibility of the judge in evaluating the reliability of scientific testimony and in appointing scientific experts as appropriate.

8. J.C. Farman, B.G. Gardiner, & J.D. Shanklin, *Large Losses of Total Ozone in Antarctica Reveal Seasonal ClO_x/NO_x Interaction*, 315 *Nature* 207–10 (1985), <https://dx.doi.org/10.1038/315207a0> [hereinafter Farman et al. 1985].

9. *Nat'l Rsch. Def. Council v. Env't Prot. Agency*, 464 F.3d 1 (D.C. Cir. 2006).

10. Oreskes & Conway 2010, *supra* note 2.

The Science Flowchart

Science is complex, iterative, dynamic, and social. It emerges out of the work of fallible people, shaped by their unique backgrounds, perspectives, and motivations, engaged in different activities in different orders as they come up with new explanations, and test and retest those ideas.¹¹ Through this process, biases are identified and corrected, inaccurate ideas are eventually rejected, and discrepancies are negotiated and resolved. This practice does lead to scientific consensus, but over many years, investigations, and interactions, not in the context of a single study. These key points are illustrated by the science flowchart (Figure 1), which more accurately represents the process of science than does the five-step scientific method frequently found in textbooks.¹²

This flowchart can help us frame scientific inquiry into a wide range of topics. It applies to the natural sciences, as well as social sciences and engineering research. It applies to both qualitative and quantitative research conducted in laboratories, in the field, in clinical settings, and through the examination of historical records and other datasets. It applies to research conducted by academics, engineers, government workers, industry employees, and experts hired by litigation teams. Any line of inquiry that aims to establish objective, reliable knowledge based on evidence can be examined using this lens. While such knowledge-building activities do not follow a prescribed step-by-step process, they *do* have some traits in common. At some point, in some way, someone gets an idea (often called a hypothesis) about how to solve a problem, explain a phenomenon, or answer a question; at some point, in some way, that hypothesis is tested against evidence; at some point, in some way, the evidence and hypothesis are scrutinized by others with relevant expertise; and if the hypothesis holds up to this scrutiny through multiple rounds of testing, people will use it—to solve problems, make decisions, and/or build more knowledge.

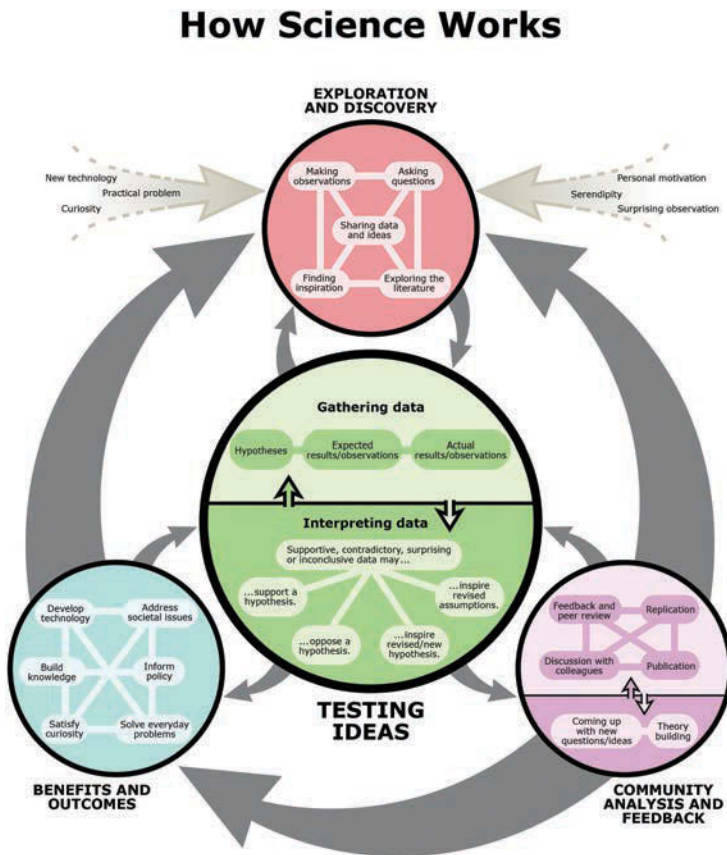
If mapped onto the science flowchart, our example of the discovery of the link between CFC production and the hole in the ozone layer would produce a winding path that circles back on itself, proceeding through the testing phase multiple times. A map of the investigations establishing that cigarettes cause lung cancer would follow a different path, as would a map of social science and psychological research examining factors that impact eyewitness recall.¹³ When a judge must evaluate

11. For example, see David L. Hull, *Science as a Process: An Evolutionary Account of the Social and Conceptual Development of Science* (1988) [hereinafter Hull 1988]; Kitcher 1995, *supra* note 5; Helen E. Longino, *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry* (1990), <https://doi.org/10.2307/j.ctvx5wbzf> [hereinafter Longino 1990]; & Wimsatt 2007, *supra* note 5.

12. University of California Museum of Paleontology, Understanding Science (2022), <https://perma.cc/MAB9-NYWL> [hereinafter UCMP 2022].

13. Each of these lines of research has, at some point, been applied in the courtroom. See Nat'l Rsch. Def. Council v. Env't Prot. Agency, 464 F.3d 1 (D.C. Cir. 2006); United States v. Philip Morris USA Inc., 566 F.3d 1095 (D.C. Cir. 2009); and Sanders v. City of Chi. Heights, No. 13 C

Figure 1. Science is complex, iterative, dynamic, and social. It does not proceed according to a linear, step-by-step process.



Source: © UC Museum of Paleontology Understanding Science, www.understandingscience.org.

whether expert scientific testimony is scientifically sound, it is not a simple matter of ensuring that a particular five-step process was followed. Instead, the judge's task requires a deeper examination of the available evidence and methods by which it was arrived at, as well as an assessment of how the community of experts in this area has evaluated or would evaluate the evidence and reasoning in question.

Widely accepted scientific knowledge is built over the course of many iterations through this flowchart—a process that often takes years. For simplicity's

0221 (N.D. Ill. Aug. 18, 2016). For more on eyewitness identification, see Thomas D. Albright & Brandon L. Garrett, *Reference Guide on Eyewitness Identification*, in this manual.

sake, judges might wish to be presented only with scientific testimony based on mature research programs, in which many studies by different groups have been conducted, evidence has been scrutinized from different viewpoints, and some level of consensus has been reached; but, of course, disputed facts in litigation may hinge on questions of science that are still emerging, and so evidence is lacking, or that are so specific that new research must be conducted to inform the litigation. Such testimony may necessarily rely on one or just a few tests that have not yet received broad scrutiny by the relevant scientific community, and it may be that the scientific investigations that would best inform litigation have not been performed; neither situation renders inadmissible testimony based on the well-conducted and informative (though not conclusive) studies that *have* been performed. Despite our reliance on an example in which scientific consensus was eventually achieved, we emphasize that scientific consensus is not a requirement for testimony to be admissible, as described in the section below and in *The Admissibility of Expert Testimony* in this manual. If two experts present conflicting scientific testimony, and the testimony for each expert is grounded in proper scientific methodology, the court may admit the testimony of both experts and allow the trier of fact (i.e., jury or judge) to resolve the conflict.

This reference guide examines the progression of science from speculation to wide acceptance, as illustrated by the CFC/ozone example, to assist with judges' evaluation of the admissibility of scientific evidence and develop their understanding of the epistemology of science.¹⁴ Subsequent reference guides will discuss the strengths and weaknesses of individual studies and lines of investigation in various scientific disciplines relevant to the judiciary.

The Science Flowchart and Admissibility

Federal Rule of Evidence 702 outlines the findings a judge must make to admit expert testimony over an objection.¹⁵ According to Rule 702, such testimony is

14. Jasanoff 2005, *supra* note 5, laments the epistemological sophistication *Daubert* would seem to require to recognize “good science,” as well as judges’ lack of training in this area—a challenge that we hope this manual will help remedy.

15. While the testimony of expert witnesses identified by the parties is the most common way that science makes its way into the courts, it is not the only way. Court-appointed experts may perform a variety of functions and services. They may serve as witnesses themselves under Federal Rule of Evidence 706; advise the court on assessment of evidence that might, for example, be relevant for admissibility decisions; help mediate settlements; and/or provide background information to the judge or jury. In litigation that involves scientific evidence, the use of such independent scientific experts may help address potential challenges to reaching a just resolution that stem from aspects of the process of science and the process of litigation highlighted herein. These include the complexities of the process of science outlined above; the importance of community analysis and

admissible when the judge finds that (1) the witness is qualified to offer the testimony, (2) the witness's expertise will help determine questions of fact, (3) the testimony is based on sufficient facts or data, (4) the testimony results from reliable principles and methods, and (5) the witness has reliably applied those principles and methods to the facts of the case. In *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, the Supreme Court found that expert scientific testimony need not be “generally accepted” to be admitted under Rule 702; instead, the Court instructed federal judges to exercise a “gatekeeping” role regarding scientific evidence, admitting for consideration only such evidence that is grounded in an appropriate “scientific methodology.”¹⁶ The Court indicated that an appropriate scientific methodology often includes the process of formulating hypotheses and conducting experiments to test the validity of the hypotheses. The Court then provided a set of illustrative criteria suggestive of the sorts of considerations judges must make in determining whether the proposed testimony meets the standard of an appropriate scientific methodology.¹⁷ The factors selected as examples by the Court include:

1. whether it can be and has been tested;
2. whether it has been subjected to peer review and publication;
3. whether it has a known error rate;
4. the existence and maintenance of standards and controls; and
5. whether the theory or technique employed by the expert is generally accepted in the scientific community.

The Ninth Circuit clarified another relevant consideration relating to *Daubert*: whether the research was conducted independently of the particular litigation or dependent on an intention to provide the proposed testimony. Subsequent

feedback in sorting through scientific explanations as described in the section titled “Science as a Human and Community Endeavor” below; the time all this takes to play out; the wide variety of valid scientific methodologies (described in the section titled “Scientific Methodologies and Data” below); the fact that individual scientists are human and inherently biased (described in the section titled “Bias” below); the depth of the subject matter knowledge that may be required to understand and evaluate scientific evidence; and of course the stakes that each party has in the testimony of their proffered expert witnesses and potential biases introduced in selecting those witnesses. Resources are available to assist judges in this process of identifying and securing a court-appointed expert—for example, the CASE program of the American Association for the Advancement of Science. For more on the services performed by court-appointed experts, see American Association for the Advancement of Science, *Court Appointed Scientific Experts: A Handbook for Experts Version 3.0* (2002), <https://perma.cc/5FHD-GKQL>, and Daniel L. Rubinfeld & Joe S. Cecil, *Scientists as Experts Serving the Court*, 147 *Daedalus* 152–63 (2018), https://doi.org/10.1162/daed_a_00526.

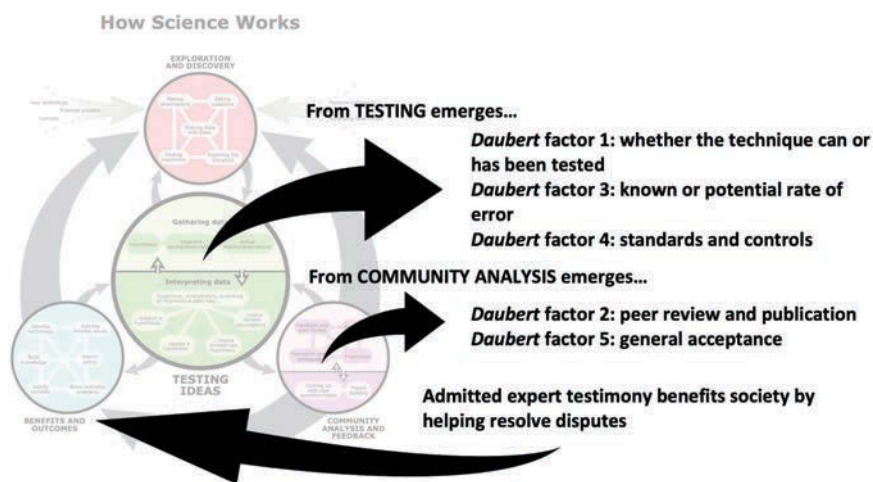
16. 509 U.S. 579 (1993).

17. For more on *Daubert* and admissibility, see Liesa L. Richter & Daniel J. Capra, *The Admissibility of Expert Testimony*, in this manual.

decisions by the Supreme Court clarified the standards for scientific and technical evidence.¹⁸

These legal standards for admissibility under *Daubert* can be mapped onto the elements of the flowchart, as illustrated by Figure 2. The central region of the flowchart outlines the logic of testing hypotheses against evidence, the first of the *Daubert* factors outlined above. Testing is essential to both the process of science and *Daubert*. If a hypothesis or testimony is not supported by or derived from evidence in some form, it cannot be based on scientific methodologies. The process of testing leads to two more of *Daubert*'s illustrative factors. The known or potential rate of error for scientific techniques or processes (the third *Daubert* factor) can be derived statistically from sufficient rounds of testing and data collection. The appropriate standards and controls required by a technique or process (the fourth *Daubert* factor) are likewise established in this way. Error rates, standards, and controls most clearly impact the consideration of studies that involve analytic and assessment techniques, such as sobriety tests, forensic identification techniques, and DNA analysis.

Figure 2. Relationships among the *Daubert* factors and the process of science.



Source: Image reproduced and adapted under an Attribution-NonCommercial-ShareAlike 4.0 (BY-NC-SA 4.0) Creative Commons License (see <https://creativecommons.org/licenses/by-nc-sa/4.0/>). Original image © UC Museum of Paleontology Understanding Science, www.understandingscience.org.

18. These standards were then codified through amendments to Federal Rule of Evidence 702. For more on Federal Rule of Evidence 702, see Liesa L. Richter & Daniel J. Capra, *The Admissibility of Expert Testimony*, in this manual.

The flowchart's *Community analysis and feedback* area highlights peer review and publication (the second *Daubert* factor) as an important element of scientific knowledge building. For many areas of science, peer-reviewed journal articles are the usual means through which the expert scientific community assesses hypotheses and evidence (see section titled "Peer Review" below). However, other scenarios are also possible. For example, an engineering technique or piece of computer code might be scrutinized through direct use by the relevant community to see if it performs as expected. Research conducted solely for litigation may be scrutinized by the judge, other expert witnesses, and a jury.¹⁹ Finally, it is through interactions within *Community analysis and feedback* and multiple iterations through the *Testing ideas* region that hypotheses, theories, and techniques gain widespread acceptance (the final *Daubert* factor; see section titled "Achieving Scientific Consensus" below). And of course, the final result of the judge's decisions on the admissibility of scientific evidence represents one of the ways that benefits (the lower left area of the flowchart) are derived from the process of science, by helping resolve disputes through litigation.

Textbook portrayals of a five-step scientific method evoke a scientist in an academic institution driven by curiosity alone.²⁰ In contrast, the upper area of the science flowchart and the *Benefits and outcomes* region illustrate the variety of factors that motivate scientific investigations. Engineering research may be driven by the need to solve a particular problem, and as *Kumho Tire Co. v. Carmichael* held, the provenance of such specialized knowledge can be meaningfully compared to standards and methodologies of good science.²¹ Career advancement, prestige, and funding are likewise incentives for many scientists.²² And as described in *The Admissibility of Expert Testimony*, in this manual, scientific research may also be conducted for the sole purpose of testifying at trial.

The fact that a study was undertaken in the service of litigation does not itself make the research unscientific or unreliable.²³ All scientists have some sort of motivation for their work, and this does not preclude scientific knowledge

19. Jasanoff 2005, *supra* note 5, describes this role of the judiciary in helping produce certain sorts of scientific knowledge, acknowledging the non-independence of science and the law.

20. McComas 1998, *supra* note 4.

21. 526 U.S. 137 (1999).

22. Hull 1988, *supra* note 11; Kitcher 1995, *supra* note 5; Longino 1990, *supra* note 11; Michael Strevens, *The Role of the Priority Rule in Science*, 100 J. of Phil. 55–79 (2003), <https://doi.org/10.5840/jphil2003100224>; and Wimsatt 2007, *supra* note 5.

23. In the 1995 remand of *Daubert* (43 F.3d 1311 (9th Cir. 1995)), Judge Alex Kozinski wrote that scientific investigations performed in the service of litigation should be more stringently evaluated for admissibility, suggesting that they may not be reliable, a perspective that has been criticized by some scholars. For example, see Sheila Jasanoff, *Representation and Re-Presentation in Litigation Science*, 116 Env't Health Persps. 123–29 (2008), <https://doi.org/10.1289/ehp.9976>; and Leslie I. Boden & David Ozonoff, *Litigation-Generated Science: Why Should We Care?* 116 Env't Health Persps. 117–22 (2008), <https://doi.org/10.1289/ehp.9987>, which are largely consistent with the broad view of science as a human and community endeavor presented in this reference guide. Indeed, Jasanoff attributes

building, so long as biased methodologies and interpretations are avoided. However, research conducted for the purpose of testifying is unlikely to have had the opportunity to be reviewed and scrutinized by the relevant scientific community. Though imperfect (see section titled “Peer Review” below), this process of community scrutiny can still help to identify and overcome biases that may shape a scientific investigation, as well as assess its methodological and deductive quality (see section titled “Science as a Human and Community Endeavor” below). Because of this, judges may be placed in the position of providing a higher level of scrutiny to testimony based on research developed for litigation than that based on established research programs conducted for other purposes, which have already iterated through the flowchart multiple times. For example, survey research is an important line of evidence in trademark infringement litigation,²⁴ and this research may be so specific in its intent (e.g., whether the term *beanies* is perceived as a brand name for a toy²⁵) that there is little reason to perform it except in service of litigation. Testimony based on such survey research is often determined to be admissible, but may not be if biases or methodological problems are identified in the study design.²⁶ In short, testimony based on research conducted solely for litigation may be scientifically reliable and admissible—or may not be; and making this determination in the absence of any of the services performed by the scientific community may place a higher burden on judges.

Key Traits of Science

Judges might wish for a single marker that distinguishes between science and pseudoscience, good and bad science, accepted and speculative science. Unfortunately, we have no easy answers. These are all the end points of continua, and multiple factors determine where on the spectrum a particular unit of knowledge or research falls. Contemporary thinking suggests that the boundary between science and non-science is best identified by the *process* by which knowledge is generated, and

this denigration of litigation-derived science to resting on “idealized, misleading, or misinformed assumptions about the scientific method,” which this reference guide aims to clarify and correct.

24. See Shari Seidman Diamond et al., *Reference Guide on Survey Research*, in this manual.

25. *Ty Inc. v. Softbelly’s, Inc.*, 353 F.3d 528 (7th Cir. 2003).

26. Rebecca Kirk Fair & Laura O’Laughlin, *Ensuring Validity and Admissibility of Consumer Surveys*, ABA Groups Practice Points (2017), <https://perma.cc/2LHD-7NE2>. For more on the design of methodologically sound surveys, see Shari Seidman Diamond et al., *Reference Guide on Survey Research*, in this manual.

especially by how the community engaged with that process behaves.²⁷ Key aspects of this process and these behaviors are outlined below.

Science Investigates the Natural World and Natural Explanations

Communities engaged in scientific endeavors aim to build accurate, natural explanations for how the world works. In this context, the term *natural* refers to any element of the physical universe—whether made by humans or not. Thus, the natural world includes matter, the forces that act on matter, energy, distant galaxies, the constituents of the biological and physical world (e.g., the ozone layer), humans, human society, and the products of that society (e.g., CFCs). As described above, the goal of building explanations does not imply that, to be scientific, research must be without practical or economic end. Both applied research, which is undertaken with the goal of solving a particular problem, and pure or basic research, which has the primary aim of building scientific knowledge, improve our understanding of the natural world.

The investigation of *supernatural* explanations—that is, explanations that rely on forces beyond the observable universe—is not a part of modern science.²⁸ The exclusion of supernatural explanations from science was recognized in *Kitzmiller v. Dover Area School District*, in which parents of school-age children brought suit against their local school board over the requirements that Intelligent Design be taught in schools and that teachers read a disclaimer that cast evolution as shaky science.²⁹ The court found that Intelligent Design is not science in part because it “violates the centuries-old ground rules of science by invoking and permitting supernatural causation.”³⁰ The philosophical and historical roots of this exclusion are deep, but a simple, practical explanation clarifies this necessity: Science is able to build reliable knowledge because it is based on evidence and observations, and supernatural explanations are generally not testable against evidence, as explained below.

27. Hull 1988, *supra* note 11; Philip Kitcher, *Believing Where We Cannot Prove*, *Abusing Science: The Case Against Creationism* 30–54 (1983) [hereinafter Kitcher 1983]; and UCMP 2022, *supra* note 12.

28. Massimo Pigliucci & Maarten Boudry, *Philosophy of Pseudoscience: Reconsidering the Demarcation Problem* (2013).

29. 400 F. Supp. 2d 707 (M.D. Pa. 2005).

30. The court’s additional reasons for finding that Intelligent Design (ID) is not science were (1) that ID proponents falsely argue that any challenge to evolutionary theory necessarily supports ID and (2) that the scientific community has investigated and rejected ID proponents’ challenges to evolutionary theory.

Science Investigates Testable Hypotheses

Communities engaged in scientific endeavors work with testable hypotheses. For a hypothesis to be testable, it must, by itself or in conjunction with other hypotheses, generate specific predictions—a set of observations that one could expect to make if the hypothesis were true and/or a set of observations that would be inconsistent with the idea and lead one to believe that it is *not* true. If an explanation is equally compatible with all possible observations, then it is not testable and hence, not within the reach of science. This is frequently the case with supernatural explanations. For example, a central proposition of Intelligent Design that a complex anatomical structure (e.g., an eye) was designed by an intelligent, supernatural entity does not make any predictions about anatomy or distribution of the structure among species.³¹ We cannot know how or why a supernatural entity would design an eye in a particular way, so *any* observation we might make about the anatomical structure is compatible with this explanation. Evidence cannot be used to investigate such propositions, and hence, they are outside the realm of science.

The logic of testing hypotheses is illustrated in the central area of Figure 1. For example, the idea that smoking causes lung cancer is testable and leads to a wide variety of predictions: that rates of lung cancer will be higher among smokers than among a comparable group of nonsmokers; that lung cancer rates will increase in populations where smoking becomes common; that exposing non-human animals to compounds in cigarettes will lead to higher rates of cancer; and many more. Similarly, the hypothesis that CFCs deplete the ozone layer led to many predictions: that certain chemical reactions occur in the atmosphere; that the distribution of CFCs and other chemicals in the atmosphere will match the predictions of models; and that the ozone layer will thin as CFCs accumulate. The same logic applies to testing hypotheses much more limited in scope. For example, the narrow hypothesis that CFCs persist in the atmosphere can be tested simply by checking for them in places far from where they were produced.³² For an example drawn from the courtroom, consider litigation surrounding whether Thermos Co.’s trademark was infringed upon by another company’s use of the term *thermos*. The narrow hypothesis that *thermos* is perceived as a generic term for an insulated vacuum bottle led to the prediction that people who want to purchase an insulated vacuum bottle from a store would ask for a “thermos.”³³

In the context of testing hypotheses, the term *prediction* refers to a logical consequence of a hypothesis, not necessarily what will happen in the future. Scientific predictions often involve reexamining records of the past. For

31. Elliott Sober, *What Is Wrong With Intelligent Design?*, 82 Q. Rev. Biology 3–8 (2007), <https://doi.org/10.1086/511656>.

32. This hypothesis was tested and described as part of Lovelock et al. 1973, *supra* note 6, although performing that test was not the main intent of the investigation.

33. King-Seeley Thermos Co. v. Aladdin Indus., 321 F.2d 577 (2d Cir. 1963).

example, key hypotheses about connections between greenhouse gases and Earth's temperature, which have dramatic implications for life on Earth today, make predictions that are best tested, in part, by examining ice and oceanic sedimentary cores that reveal Earth's deep history.³⁴

Science Responds to Evidence

Science builds knowledge based on evidence; this is how the community ensures that the explanations it generates approach accuracy. As a community, scientists actively seek evidence to test their hypotheses, and ultimately accept, give up, or modify hypotheses as warranted by the evidence. Just as Rule 702(d) directs courts to assess whether there is a disconnect between an expert's methods and principles, and the conclusions reached, reliable scientific knowledge building depends on there being a logical connection between what the evidence shows and what explanatory hypotheses are accepted. Ignoring evidence is not an option in science. For example, when research revealed the hole in the ozone layer to be much larger than any of the models predicted, scientists rushed to explore additional explanations and modifications of the key hypothesis that would help make sense of this new evidence.³⁵

Evidence is not the *only* arbiter of whether a hypothesis is accepted by the scientific community. Other considerations, like broadness and coherence, also matter. Furthermore, the community's response to evidence or engagement with testing a hypothesis may be slower or hindered if dogma or other social factors work against it, as described in the section titled "Achieving Scientific Consensus" below. But evidence is the most important determinant of acceptance in science. Explanations that are not supported by evidence are eventually rejected. Hypotheses that are protected from testing or that are allowed to be tested by only one group of investigators with a vested interest in the outcome are not a part of good science.

Science Does Not *Prove* Hypotheses

Because science responds to evidence, scientific knowledge is always open to refinement, revision, or even rejection if warranted. Philosophers of science call this property *fallibilism*.³⁶ No matter how much evidence supports or refutes it, a

34. Valérie Masson-Delmotte et al., *Information from Paleoclimate Archives*, in Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change 383–464 (2013).

35. E.g., Susan Solomon et al., *On the Depletion of Antarctic Ozone*, 321 Nature 755–58 (1986), <https://doi.org/10.1038/321755a0>.

36. Susan Haack, *Inquiry and Advocacy, Fallibilism and Finality: Culture and Inference in Science and the Law*, 2 L., Probability & Risk 205–14 (2003), <https://doi.org/10.1093/lpr/2.3.205>.

hypothesis cannot be absolutely proven true or false.³⁷ Even widely accepted scientific explanations, like the link between CFCs and ozone depletion, could be rejected if new evidence or a better-supported hypothesis comes along. This has happened at many points in the history of science. For example, Newton's classical mechanics is a powerful predictive tool that was relied upon by physicists for 200 years and is still used to generate reliable predictions in many situations. But when Einstein's theory of relativity showed itself to be more closely aligned with the evidence in more situations than Newtonian mechanics, scientists accepted that this new framework was a broader and more powerful one. At a smaller scale, the idea that one misfolded protein can cause others to misfold and generate disease (i.e., the prion hypothesis) was once regarded as beyond far-fetched and rejected by biologists; however, as evidence accrued showing prions to be proteins responsible for transmissible diseases such as mad cow disease and scrapie, dogma was overridden and scientists accepted this expanded conception of proteins and disease.³⁸ As new scientific facts are established and insufficient ones are overturned, judges can expect these slow-moving ripples of change to play out at trial. For example, bitemark evidence may be increasingly excluded as research establishes its fundamental limitations.³⁹

Despite its tentative nature, accepted scientific knowledge is reliable. Such explanations generate predictions that hold true in many different contexts and at many different scales, allowing us to figure out how entities in the natural world are likely to behave and how we can harness that understanding to solve problems and dispense justice. For example, Molina and Rowland's understanding of chemistry allowed them to predict that CFCs were damaging the ozone layer long before this was observed. Because of this scientific knowledge, we were able to make policy changes and ultimately reverse damage to the ozone layer. We have good reason to trust accepted scientific explanations: They allow us to make accurate predictions about the future and intervene on the present. Even Newtonian mechanics, though subsumed by relativity at an explanatory level, is still used at a practical level for a wide variety of applications, such as building bridges and predicting the movement of satellites and asteroids.

37. Formal proofs are central to mathematics, a field that provides critical tools for scientists but that is nonetheless distinct from scientific disciplines that rely on evidence from the natural world to assess hypotheses. Mathematical proofs accept a set of axioms as true and demonstrate that a particular conclusion is logically implied by those assumptions. While the sciences described here incorporate logic and inference, their reliance on evidence and observations to verify conclusions means that hypotheses are fallible in a way that mathematical statements are not.

38. Claudio Soto, *Prion Hypothesis: The End of the Controversy?*, 36 *Trends in Biochemical Scis.* 151–58 (2011), <https://doi.org/10.1016/j.tibs.2010.11.001> [hereinafter Soto 2011].

39. See Valena E. Beety et al., *Reference Guide on Forensic Feature Comparison Evidence*, in this manual, for more on bitemark evidence.

Science Is Carried Out by a Community that Holds Members to Norms

Communities engaged in scientific endeavors generally adhere to a set of practical behavioral standards, known as norms, which can change over time. There is no one group tasked with ensuring adherence to scientific norms, but because they are recognized as important by a wide variety of scientific institutions, scientists who are found to violate them knowingly are likely to be impeded from publishing, receiving grants, and advancing their career (see section titled “Misconduct” below for more on this). The following list summarizes some important, current scientific norms:

- *Seek relevant information:* Because science is cumulative and scientists are in the business of seeking broad explanations, scientists need to know what others before them have learned about a system to form hypotheses that cohere with the many lines of available evidence. Science rarely starts from scratch. For example, upon hearing how widespread CFCs were, Mario Molina first turned to the published work of other scientists to see what chemical processes might break down CFCs in the lower atmosphere before they reach the ozone layer.
- *Scrutinize ideas and evidence:* In science, questioning methods and looking for other explanations for evidence does not necessarily signal that a hypothesis is wrong or a study was poorly executed. For example, the prion hypothesis was once viewed suspiciously by many biologists, who required particularly convincing evidence to accept an idea that seemed at odds with how proteins were known to behave.⁴⁰ Such skepticism is part of the normal practice of science and helps steer it in the direction of accuracy. Indeed, scientists are generally motivated to scrutinize closely even well-established ideas and evidence because of the esteem that the community awards those who overturn accepted science and contribute truly new and fruitful explanations.
- *Openly communicate findings:* Scientists are expected to share their hypotheses, methods, and findings honestly and openly to allow others to evaluate and build on them. When Rowland and Molina uncovered evidence about the longevity of chlorine nitrate in the atmosphere that threatened their hypothesis, they published the evidence for others to evaluate.⁴¹ For examples of how violation of this norm can lead to or factor into litigation, see section titled “Misconduct” below.

40. Soto 2011, *supra* note 38.

41. F. Sherwood Rowland et al., *Stratospheric Formation and Photolysis of Chlorine Nitrate*, 80 J. Physical Chemistry 2711–13 (1976) [hereinafter Rowland et al. 1976].

- *Identify and avoid bias:* Because scientists are human, it is impossible for them to be completely objective and unbiased in their work (see section titled “Science as a Human and Community Endeavor” below). However, scientific progress depends on scientists aspiring toward this ideal by *trying* to employ fair tests and methods, and to evaluate ideas on the merits of the evidence. For example, the scientists who first reported the hole in the ozone layer over Antarctica strove for objectivity in several ways. They got their data from carefully calibrated instruments, directed other scientists to information about how the machines were calibrated, and presented data with generous error bars to accommodate subsequent corrections to their observations.
- *Assimilate evidence:* When faced with evidence contradicting their hypothesis, scientists may decide to conduct more tests, may revise or reject the idea, or may consider alternative ways to explain the evidence. But ultimately, scientific explanations are sustained by evidence and cannot be propped up—or denied—if the evidence is clear. For example, as evidence mounted in favor of the prion hypothesis, even skeptics were convinced.
- *Provide appropriate credit:* Scientists are expected to cite their sources scrupulously. Credit is an important measure of impact within the scientific community. It also provides a sort of paper trail, through which another scientist can evaluate the methods, assumptions, and reasoning that a particular study is based on. Rowland and Molina’s original paper on the CFC/ozone link included 31 citations outlining where their ideas and evidence originated.⁴²
- *Abide by official ethical guidelines:* Research institutions, funding agencies, publishers, scientific societies, and legislation regulate what scientific research can be done and how. In general, these guidelines are intended to ensure that scientific work is of high quality, is performed in ethical ways, and benefits society. Such guidelines vary depending on the subject, application, scope, and setting of research. For example, a survey-based study designed to contribute to generalizable knowledge about human perception will generally need to be approved by an institutional review board (IRB) tasked with protecting the interests and rights of survey participants before the research can commence.

When the work of scientific experts aligns with these norms, judges can be more confident that their testimony stems from reliable scientific methodologies.

42. Molina & Rowland 1974, *supra* note 7.

Nonscience, Pseudoscience, Bad Science, and Wrong Science

The characteristics above can help us identify research that is clearly scientific, but there are different ways for an endeavor to be *unscientific*. Some lines of inquiry are simply outside the realm of science and do not pretend to be otherwise. For example, science cannot make moral judgments. Science can help us assess what segment of the population finds euthanasia acceptable, understand how people reason about euthanasia as they near the end of their lives, and learn about the emotional effects of euthanasia,⁴³ but science cannot tell us if euthanasia is morally right or wrong.

Pseudoscience, on the other hand, employs the trappings of science but fails to meet several key scientific standards. For example, homeopathy is a type of alternative medicine that arose in the 18th century out of legitimate concerns about the efficacy of mainstream medicine at the time. It is based on the idea that “like cures like”—that ingesting a substance that causes certain bodily effects in a healthy person (e.g., chills) can be used to cure a disease that causes that symptom. Homeopathy uses Latin scientific names for the substances in its treatments, has adopted the vocabulary of mainstream medicine (e.g., *vaccine*), and often packages and labels products in ways that are reminiscent of mainstream pharmaceuticals—all of which might be interpreted as somehow scientific. However, many evidence-based tests of homeopathic remedies have been carried out and turned up no data suggesting that they perform better than placebos. Similarly, no evidence of a natural mechanism that might underlie the claims made by homeopaths has been found. Many panels, committees, and councils tasked with assessing the scientific consensus on homeopathy have concluded that it has no scientific basis and is, instead, pseudoscience.⁴⁴ In *Allen v. Hyland’s, Inc.*, the plaintiffs argued that Hyland’s made false claims about the efficacy of its homeopathic remedies.⁴⁵ The defendants’ expert witnesses were allowed to testify, a decision that has been criticized as a failure of the court’s scientific gatekeeping function

43. For example, see Joachim Cohen et al., *European Public Acceptance of Euthanasia: Socio-Demographic and Cultural Factors Associated with the Acceptance of Euthanasia in 33 European Countries*, 63 Soc. Sci. & Med. 743 (2006), <https://doi.org/10.1016/j.socscimed.2006.01.026>; Ronit D. Leichentritt & Kathryn D. Rettig, *Meanings and Attitudes Toward End-of-life Preferences in Israel*, 23 Death Stud. 323–58 (1999), <https://doi.org/10.1080/074811899200993>; and Nikkie B. Swarte et al., *Effects of Euthanasia on the Bereaved Family and Friends: A Cross Sectional Study*, 327 BMJ 189 (2003), <https://doi.org/10.1136/bmj.327.7408.189>.

44. For example, see National Health and Medical Research Council, *NHMRC Information Paper: Evidence on the Effectiveness of Homeopathy for Treating Health Conditions* (2015).

45. No. CV 12–1150 DMG (MANx) (C.D. Cal. Aug. 16, 2016); See Robert G. Knaier, *Homeopathy on Trial: Allen v. Hyland’s, Inc. and a Failure of Evidentiary Gatekeeping*, 57 Jurimetrics J. 361–96 (2017).

under *Daubert*, and the jury found for the defendants, highlighting the challenges of detecting pseudoscience and assessing scientific consensus.

Then there is “bad” science: specific investigations that are within the realm of science (focus on natural explanations, work with testable ideas, etc.) but are unreliable because they have not been executed in ways that meet the community’s standards for scientific behavior—perhaps because the researchers selectively reported data and so did not openly and honestly communicate findings. Some analyses suggest that cases of bad science are on the rise because of systemic incentives that reward quantity of publications and certain types of results over quality.⁴⁶ A fuller framing of this concern is provided in the section below titled “Replication and the ‘Replication Crisis.’”

Unsurprisingly, bad science often leads to wrong science—that is, acceptance of hypotheses that are inaccurate. Even investigations that meet all the standards described above can wind up supporting incorrect explanations. But because science involves retesting hypotheses in different ways, contexts, and combinations, inaccurate ideas will ultimately be identified as such. This is part of the normal way in which science works, but it does take time, often many years, for scientists to sort through explanations and identify those that consistently lead to accurate predictions, new insights, and improved understanding.

Sometimes entire fields or lines of inquiry are denounced as not scientific or pseudoscientific, and such arguments may be leveraged at trial. However, as we’ve seen, science can address a vast array of topics. A finer-grained analysis may be required to assess *which* aspects of a body of knowledge are backed by scientific evidence and which are not. For example, acupuncture is a technique of traditional Chinese medicine used for more than 2,000 years that is based on the idea of Qi (or “Chi”), a type of life-giving energy that circulates through the body in special channels. No scientific evidence supports the existence of Qi or the channels through which it circulates. If we accept the traditional view that Qi is a mystical force, then this explanatory mechanism is *supernatural* and outside the realm of science. However, sound scientific investigations have found acupuncture to be useful for treating a variety of symptoms, and scientific research is beginning to investigate the biological basis of these effects.⁴⁷ At the same time, proponents sometimes overstate these findings, and practitioners may employ terminology and tools that patients associate with medicine (often for good reason), making it difficult to distinguish which aspects of acupuncture are supported by scientific research and which are pseudoscientific. However, labeling *all* of acupuncture as nonscience ignores reliable scientific evidence and knowledge.

46. For an overview of the concern, causes, and proposed solutions, see National Academies of Sciences, Engineering, and Medicine, *Reproducibility and Replicability in Science* (2019) [hereinafter NASEM].

47. For example, see Tony Y. Chon & Mark C. Lee, *Acupuncture*, 88 Mayo Clinic Proc. 1141–46 (2013), <https://doi.org/10.1016/j.mayocp.2013.06.009>.

Science as a Human and Community Endeavor

Science is carried out not by objective automatons but by people, shaped by their unique backgrounds and motivations, as well as by the changing norms and values of the societies in which they are embedded, all of which influence the course of science. The identities, backgrounds, and experiences of the practitioners of science impact their selection of research topics, hypotheses, chosen research methods, and interpretations of results and evidence. Such influences are apparent in a wide range of scientific fields but are famously exemplified by shifts in the field of primatology as more women entered the field in the 1970s. These scientists did not make the same assumptions that their male predecessors had made and so observed and investigated behaviors and interactions previously overlooked and undocumented, revealing, for example, that female primates have elaborate sex lives and manipulate male behavior.⁴⁸ Of course, the fact that scientists are fallible humans also means that they are vulnerable to bias and capable of misconduct, as described in the sections below.

The human practitioners and institutions of science likewise respond to events in and values of society around them in various ways. For example, the 1970s energy crisis led to a surge of alternative energy research and reorganizations of funding mechanisms for this research.⁴⁹ To cite an even broader

48. For a few examples of how scientists' identities, backgrounds, and experiences impact their science, across several disciplines, including in the field of primatology, see Donna Haraway, *Primate Visions: Gender, Race, and Nature in the World of Modern Science* (1989); Andrew Gary Darwin Holmes, *Research Positionality—A Consideration of Its Influence and Place in Qualitative Research—A New Researcher Guide*, 8 *Shanlax Int'l J. of Educ.* 1–10 (2020), <https://doi.org/10.34293/education.v8i4.3232>; Rembrand Koning, Sampsa Samila, & John-Paul Ferguson, *Who Do We Invent For? Patents by Women Focus More on Women's Health, but Few Women Get to Invent*, 372 *Science* 1345–48 (2021), <https://doi.org/10.1126/science.aba6990>; Diego Kozlowski et al., *Intersectional Inequalities in Science*, 119 *Proc. of the Nat'l Acad. of Sci. USA* e2113067119 (2022), <https://doi.org/10.1073/pnas.2113067119> [hereinafter Kozlowski et al., 2022]; D. N. Lee, *Taking it Personally*, 311 *Sci. Am.* 47 (2014), <https://doi.org/10.1038/scientificamerican1014-47>; Douglas Medin, Carol D. Lee, & Megan Bang, *Particular Points of View*, 311 *Sci. Am.* 44–45 (2014), <https://doi.org/10.1038/scientificamerican1014-44>; H. Richard Milner, IV, *Race, Culture, and Researcher Positionality: Working through Dangers Seen, Unseen, and Unforeseen*, 36 *Educ. Researcher* 388–400 (2007), <https://doi.org/10.3102/0013189X07309471>; Mathias Wullum Nielsen et al., *One and a Half Million Medical Papers Reveal a Link Between Author Gender and Attention to Gender and Sex Analysis*, 1 *Nature Hum. Behav.* 791–96 (2017), <https://doi.org/10.1038/s41562-017-0235-x>; Cassidy R. Sugimoto et al., *Factors Affecting Sex-Related Reporting in Medical Research: A Cross-Disciplinary Bibliometric Analysis*, 393 *Lancet* 550–59 (2019), [https://doi.org/10.1016/S0140-6736\(18\)32995-7](https://doi.org/10.1016/S0140-6736(18)32995-7); Caitlin Wilson, Gillian Janes, & Julia Williams, *Identity, Positionality and Reflexivity: Relevance and Application to Research Paramedics*, 7 *Brit. Paramedic J.* 43–49 (2022), <https://doi.org/10.29045/14784726.2022.09.7.2.43>.

49. Alice L. Buck, *History of the Energy Research and Development Administration*, No. DOE/ES-0001 USDOE Assistant Secretary for Management and Administration (1982).

example, while the earliest examples of scientific peer review date back hundreds of years, scholarship has traced the rise of peer review as a central pillar of scientific publication and funding—as well as its perception as a critical hallmark of “good” science—to the Cold War. That conflict instigated an expansion of public monetary support for scientific research and, as society’s anxieties about American scientific competitiveness began to ease, led to a tension between public accountability and scientists’ personal desires for autonomy, respect, and financial support.⁵⁰ Individual scientists and the overarching institutions of science are shaped by society and, in turn, produce scientific knowledge that goes on to affect society through its applications and implications, creating the feedback loop between “benefits and outcomes” and other parts of the process of science illustrated in Figure 1. Both scientific institutions and scientists are deeply embedded within and interconnected with broader society.

Social interactions within the scientific community are critical to building scientific knowledge. We saw this in the CFC/ozone example as one researcher’s conference presentation sparked Molina and Rowland’s first investigation of CFCs, which, in turn, led to a cascade of studies by different research groups. Scientists often work collaboratively, and today more and more science is done by large research groups and multi-institution consortia, reflecting the increasing complexity of scientific questions being investigated and the specialized knowledge needed to carry out those investigations. The scientific community performs other key services as well. One of the most important of these is scrutinizing the evidence and ideas of other scientists.

Scientific scrutiny often involves the following activities, which help ensure the reliability of the scientific knowledge produced:

- an evaluation of methods, considering potential biases and oversights,
- reconsideration of the hypothesis and how to interpret the evidence relevant to it, and
- appraisal of alternative hypotheses that may also explain the results.

Scientists engage in these activities at many times, including when serving as peer reviewers for a publication, when reviewing research for their own information, and through questions and discussions at conferences. In their published articles, scientists generally preserve an account of how they have applied this scrutiny to others’ work and to their own research described in the article. For example, a key paper that helped show that proteins are the infectious agents in prion diseases documents the careful and often laborious reasoning required

50. Melinda Baldwin, *Scientific Autonomy, Public Accountability, and the Rise of “Peer Review” in the Cold War United States*, 109 *Isis* 538–58 (2018) [hereinafter Baldwin 2018].

to fully scrutinize a hypothesis and the relevant evidence.⁵¹ In that paper, the author critiques other researchers' methods of purifying the infectious agent and explains why they might not be useful; considers fourteen different previously proposed alternative hypotheses regarding the nature of the infectious agent in prion diseases; outlines the many different lines of evidence relating to these hypotheses; considers modifications of these hypotheses that might still account for the available data; describes where the author's findings differ from those of other investigators; explains how one set of data the author collected is consistent with multiple hypotheses; acknowledges when clearly relevant factors could not be taken into account in the author's calculations; and describes where more research needs to be done because of known problems with certain assays. The application of such a high degree of scrutiny reflects the goal of producing accurate explanations for the natural world despite the fundamentally fallible nature of scientific knowledge. When research relevant to a trial has not yet been scrutinized by a community with the appropriate technical expertise, a judge may be placed in the position of providing or requesting this scrutiny.

Ultimately, through such scrutiny, scientists decide what ideas, results, and methods are worth building on and retesting in the same or some other context. Because scientists are unique, societally situated, and fallible humans, the scientific community can better perform the functions listed above, as well as many others, when scientists represent the diversity of the societies in which science is embedded.⁵² A diverse scientific community balances biases, investigates areas of inquiry relevant to all parts of society, and explores more innovative ideas for addressing the challenges it sets for itself.⁵³ Science is making progress toward equitable inclusion, albeit slowly.⁵⁴

51. Stanley B. Prusiner, *Novel Proteinaceous Infectious Particles Cause Scrapie*, 216 *Science* 136–44 (1982), <https://doi.org/10.1126/science.6801762>.

52. Here, the term *diversity* includes racial and gender identity, of course, but also many other facets of identity and background, including culture, religion, age, sexual orientation, disability, incarceration history, class, and more.

53. For evidence that diversity fosters the investigation of diverse questions of relevance to society, see Travis A. Hoppe et al., *Topic Choice Contributes to the Lower Rate of NIH Awards to African-American/Black Scientists*, 5 *Sci. Advances* 1–12 (2019), <https://doi.org/10.2226/sciadv.aaw7238>, and Kozlowski et al., 2022, *supra* note 48. For evidence that diversity fosters innovation, see Bas Hofstra et al., *The Diversity-Innovation Paradox in Science*, 117 *PNAS* 9284–91 (2020), <https://doi.org/10.1073/pnas.1915378117>. For more on the mechanisms behind this connection, see Patrick Grim et al., *Diversity, Ability, and Expertise in Epistemic Communities*, 86 *Phil. Sci.* 98–123 (2019), <https://doi.org/10.1086/701070>.

54. For more on changes in the diversity of scientific workforce in the U.S., see Pew Rsch. Ctr. *STEM Jobs See Uneven Progress in Increasing Gender, Racial, and Ethnic Diversity* (2021). For a global picture, see Fred Guterl, *Where Are the Data?*, 311 *Sci. Am.* 40–41 (2014).

Peer Review

Peer-reviewed journal articles are the most important way that scientists convey their ideas and evidence to the scientific community, offering them up for scrutiny. Indeed, in *Kitzmiller v. Dover Area School District*, the court found that Intelligent Design could not be considered an accepted scientific theory in part because its proponents had not published research in peer-reviewed journals. That is not to suggest that all testimony at trial must be based on peer-reviewed research, but that information required to be addressed in the science classroom should meet this minimum standard. *Daubert*, of course, includes peer review as one of five nonexclusive factors judges may consider in assessing the scientific validity of expert testimony.

In the process of peer review, editors at a journal, typically themselves scientists, receive a submission and first decide if it is appropriate for publication in that journal: whether it is sound enough, on topic, and of sufficient importance. If they determine that it is, they send it to other scientists with expertise on the topic of the article—the peer reviewers—who are not known to have a conflict of interest: for example, scientists who are not direct collaborators of the paper’s authors or identified by the authors as potentially vested in the publication decision. The peer reviewers may recommend rejection or acceptance of the paper outright or may request changes, starting a back-and-forth among the authors, reviewers, and editors until the editors make a final decision on acceptance or rejection. Peer review and publication are time-consuming, usually involving many months between submission and publication. The process can also be highly competitive. The journal *Science* accepted just 6.4% of the articles it received in 2021.

When the institution of peer review was adopted by the scientific community, it was not intended to distinguish credible science from that which is inaccurate or badly executed, though the process is now often viewed as serving that function.⁵⁵ Nevertheless, peer review is an important gatekeeper to the official record of science, ensuring that published works at least purport to meet minimum standards of quality and objectivity. In this way, it is reasonable to consider whether scientific knowledge that might impact a trial has been published in a peer-reviewed journal. However, peer-reviewed publication is a far from sure-fire indicator of reliable science. Here are some of the reasons.

- Peer-reviewed journals vary in their standards and criteria for publication. Some are much more lax about methodological rigor than others, and often these differences are readily apparent only to members of the expert scientific communities themselves. Judges may be tempted to rely on a journal’s impact factor, a statistic that indicates how often articles

55. Baldwin 2018, *supra* note 50.

published in that journal are cited and hence loosely conveys its prestige within a field, as an indication of methodological rigor. However, prestige is not a simple corollary of quality or correctness. A journal may have high standards for publication but a lower impact factor because of its topic or the type of articles it publishes (e.g., review articles are likely to garner more citations). Furthermore, journals can use various strategies to artificially boost their impact factor.⁵⁶ To understand the quality of a particular specialized journal, a judge may need to rely on experts within the field.

- Most but not all of the material that is published in peer-reviewed journals is original research. In addition to research articles, such journals publish commentary, including editorials, that may or may not be peer reviewed and that may express opinions and perspectives not immediately derived from evidence, and identifying these commentaries as such can sometimes present a challenge to nonspecialists.
- Recognizing the weight that peer-reviewed publications are given, sometimes regardless of the quality of the evidence they present, proponents of pseudoscientific or rejected lines of inquiry have begun to put out peer-reviewed journals that are oriented toward promoting a particular viewpoint, not objectively weighing different lines of evidence.
- Peer reviewers usually cannot detect scientific fraud. For example, widely cited research investigating the cause of Alzheimer’s disease published in the prestigious peer-reviewed journal *Nature* in 2006 allegedly includes fraudulent, manipulated images.⁵⁷ When scientific fraud is discovered, it may result in retraction (see section titled “Correction and Retraction” below), but this process of discovery and resolution takes time. The validity of the 2006 Alzheimer’s disease paper’s results was not brought into question until 2021, when an unaffiliated scientist was commissioned to investigate the research by an attorney for investors.⁵⁸ Then in 2022, *Nature*’s editors attached a note to the publication cautioning readers that concerns had been raised and that an investigation was ongoing. The paper was retracted on June 24, 2024.
- Many investigations that appear in peer-reviewed journals may reach an incorrect conclusion, which is revealed only when further testing is applied to the hypothesis. This is a normal part of the process of science.

56. Douglas N. Arnold & Kristine K. Fowler, *Nefarious Numbers*, 58 Notices Am. Mathematical Soc’y 434–37 (2011), <https://doi.org/10.48550/arXiv.1010.0278>.

57. Sylvain Lesné et al., *A Specific Amyloid- β Protein Assembly in the Brain Impairs Memory*, 440 *Nature* 352–57 (2006), <https://doi.org/10.1038/nature04533> [hereinafter Lesné et al. 2006].

58. Charles Piller, *Blots on a Field?*, 377 *Science* 358–63 (2022), <https://doi.org/10.1126/science.add9993>.

- Despite efforts to eliminate it, bias may influence the peer review of individual papers, as well as editors' decisions about publishing those papers. Some have argued that the process may be conservative and favor established hypotheses and less innovative approaches.⁵⁹ Others have suggested that the process currently favors certain types of positive results, even if they turn out to be incorrect.⁶⁰ And studies have reported different findings in terms of whether historically oppressed groups experience bias in the peer-review process.⁶¹

Rowland and Molina's CFC hypothesis was first published in *Nature*, which aims to publish original work that is of widespread interest and importance across science. Rowland and Molina's work met both of these requirements and ultimately led to a widely accepted scientific hypothesis. One might think that if a journal like *Nature* publishes a paper, its hypotheses are more likely correct, at least in comparison with hypotheses published in less-competitive journals. However, because such journals require that the work they publish be of widespread interest and importance, in fact, the papers that appear in them may be more likely to concern "risky" hypotheses that make bold explanations challenging established ideas or that leverage new techniques still undergoing development. In addition, "everyday" science, which applies widely accepted methods in routine ways, is unlikely to warrant publication in preeminent journals. Journal prestige is thus not a clear indicator of scientific reliability. We saw that Rowland and Molina's hypothesis was basically correct when it was published in *Nature*, but it still underwent many refinements and modifications afterward, particularly to the mechanisms involved in CFCs' depletion of the ozone layer.

Some new research may be on track to undergo peer review but be so recent that this has not yet taken place. Such research is often shared in the form of preprints, fully drafted articles that include all the features of scientific journal articles but have not yet been accepted to a peer-reviewed journal for publication.⁶² Preprints are often shared via online repositories (often with names of the format arXiv, bioRxiv, medRxiv, etc.). Such repositories may accept or

59. Kyle Siler, Kirby Lee, & Lisa Bero, *Measuring the Effectiveness of Scientific Gatekeeping*, 112 PNAS 360–65 (2014), <https://doi.org/10.1073/pnas.1418218112>.

60. For example, see Paul E. Smaldino & Richard McElreath, *The Natural Selection of Bad Science*, 3 Royal Soc'y Open Sci. 160384 (2016), <https://dx.doi.org/10.1098/rsos.160384>.

61. For a few examples, see Donna K. Ginther et al., *Race, Ethnicity, and NIH Research Awards*, 333 Science 1015–19 (2012), <https://doi.org/10.1126/science.119783>; Markus Helmer et al., *Gender Bias in Scholarly Peer Review*, 6 eLife e21718 (2017), <https://doi.org/10.7443/eLife.21718>; Patrick S. Forscher et al., *Little Race or Gender Bias in an Experiment of Initial Review of NIH R01 Grant Proposals*, 3 Nature Hum. Behav. 257–64 (2019), <https://doi.org/10.1038/s41562-018-0517-y>; and Flaminio Squazzoni et al., *Peer Review and Gender Bias: A Study on 145 Scholarly Journals*, 7 Sci. Advances 1–11 (2021), <https://doi.org/10.1126/sciadv.abd0299>.

62. See Oya Y. Rieger, Preprints in the Spotlight: Establishing Best Practices, Building Trust (2020), <https://doi.org/10.18665/sr.313288>.

reject submissions, but the process behind these determinations may not be disclosed and is not standardized in the way that peer review typically is. Most investigators who submit to preprint repositories *intend* for the work to undergo peer review and publication eventually but may not be successful at that. The fact that an article is called a *preprint* does not mean that publication is imminent. Preprints are beginning to make their way into the courts. For example, in a recent case, an incarcerated person moved to be granted compassionate release based on the risks he faced if he were to contract Covid-19.⁶³ The court denied the motion in part because the plaintiff had been vaccinated, mitigating his risk, and cited several lines of evidence supporting this view, including a preprint examining the efficacy of the Moderna vaccine against different strains of SARS-CoV-2.⁶⁴

Furthermore, some technical research relevant to trials, particularly that undertaken for the purpose of litigation and that conducted within industry to support functions of narrow interest, will not have had the opportunity to undergo peer review because of its provenance and intent. As addressed in *The Admissibility of Expert Testimony* in this manual, the fact that research has not undergone peer review does not make testimony based on it inadmissible; however, it may place a higher burden on the experts to explain their methods—and on the judge to assess the validity of this explanation.

Bias

As described above, scientists hold each other to a set of norms that help science build accurate explanations. However, scientists are human beings and, while they strive toward objectivity, they also bring their unique backgrounds and experiences to bear on their work. This invigorates problem solving, sparks new ideas, and benefits science in many ways. But it also means that the unconscious and conscious biases we all hold can influence the course of science and that scientists may interpret the same data in different ways.⁶⁵ The norms of scientific behavior, for example the expectations of scrutiny and open communication, as well as processes involving the scientific community like the institution of peer review, serve the function of helping identify and correct biases. Scientific objectivity thus lies in a diverse community of inquirers, not individual scientists.

63. United States v. Singh, 525 F. Supp. 3d 543 (M.D. Pa. 2021).

64. Kai Wu et al., *mRNA-1273 Vaccine Induces Neutralizing Antibodies Against Spike Mutants from Global SARS-CoV-2 Variants*, bioRxiv <https://perma.cc/SU3B-9ZCZ>.

65. Heather E. Douglas, *Science, Policy, and the Value-Free Ideal* (2009), <https://doi.org/10.2307/j.ctt6wrc78>; Hugh Lacey, *Is Science Value Free?* (2004), <https://doi.org/10.4324/9780203983195>; and Miriam Solomon, *Social Empiricism* (2007).

In examining the methodologies and conclusions of expert testimony, the source of funding for the research is a valid consideration, though not in and of itself a reason to exclude testimony. Much evidence suggests that funding source is correlated with study outcome.⁶⁶ For example, in pharmaceuticals, there is a strong tendency for industry-sponsored trials to favor the industry's product.⁶⁷ Several possible, non-mutually exclusive explanations may underlie this general effect, and they are not all nefarious—e.g., perhaps drug companies mainly fund clinical trials when internal research already strongly supports a compound's efficacy. However, biased study design or biased interpretation directly introduced by the funder or by a potentially unconscious sense of quid pro quo on the part of the researcher is also a possibility.

Many potential sources of bias can be subtle and/or hard to detect. Within this manual, the reference guides on forensic science, statistics and research methods, survey research, and epidemiology outline many ways that methods and analytic techniques can subtly bias research outcomes and, hence, testimony. And of course, research sponsored by a defendant or plaintiff that fails to support the sponsor's interest or that turns out to support the interests of the other party may simply not be introduced at trial.⁶⁸ In other cases, the influence of sponsorship is obvious. For example, after Molina and Rowland's hypothesis began to gain traction, a CFC-industry-backed group sponsored a speaking tour by a meteorologist who considered the hypothesis "utter nonsense"—an approach that seemed to be recognized by the media for what it was, a public relations stunt.⁶⁹

That is not to suggest that government- or nongovernmental organization (NGO)-sponsored research is necessarily free from bias. Individuals, academic

66. For reviews, see Sheldon Krinsky, *Do Financial Conflicts of Interest Bias Research? An Inquiry into the "Funding Effect" Hypothesis*, 38 Sci., Tech., & Hum. Values, 566–87 (2012), <https://doi.org/10.1177/0162243912456271>; and David B. Resnik & Kevin C. Elliott, *Taking Financial Relationships into Account When Assessing Research*, 20 Accountability Rsch. Policies & Quality Assurance, 184–205 (2013), <https://doi.org/10.1080/08989621.2013.788383>.

67. Sergio Sismondo, *Pharmaceutical Company Funding and Its Consequences: A Qualitative Systematic Review*, 29 Contemp. Clinical Trials 109–13 (2008), <https://doi.org/10.1016/j.cct.2007.08.001>.

68. On the other hand, failing to conduct or report on a study that a litigant could have conducted may be viewed with suspicion, as noted in *Eagle Snacks, Inc. v. Nabisco Brands, Inc.*, where the court observed that "failure of a trademark owner to run a survey to support its claims of brand significance and/or likelihood of confusion, where it has the financial means of doing so, may give rise to the inference that the contents of the survey would be unfavorable. . . ." 625 F. Supp. 571 (D.N.J. 1985).

69. The Council on Atmospheric Sciences sponsored the tour by British scientist Richard Scorer. Lydia Dotto & Harold Schiff, *The Ozone War* (1978) [hereinafter Dotto & Schiff 1978]; Edward A. Parson, *Protecting the Ozone Layer: Science and Strategy* (2003) [hereinafter Parson 2003]; and Walter Sullivan, *Scientist Doubts Spray Cans Imperil Ozone Layer*, N.Y. Times (July 8, 1975).

institutions, and other organizations may also have conflicts of interest (COIs) that are not immediately apparent—for example, through stock ownership. Government-sponsored researchers may feel motivated to secure their next grant with a result that will impress or that aligns with an agency’s interests or policy positions. In an example that goes beyond mere bias, the potentially fraudulent Alzheimer’s research described above was funded by a government agency, the National Institutes of Health, and was carried out by researchers mainly associated with universities.⁷⁰ Nevertheless, at the time of publication of that paper, two of the authors were listed as inventors on a patent application stemming from the research submitted by the public university that employs them. Clearly, receiving sponsorship from a government agency and being situated in an academic setting is no guarantee that research is untainted. All research is potentially influenced by bias, and *every* funder of research has the potential to introduce a source of bias. Evidence has highlighted the strength of this connection when it comes to industry-sponsored research. For this reason, disclosure of funding sources and COIs is generally a requirement of peer-reviewed publication. Molina and Rowland, for example, noted in their original paper that it was funded by the U.S. Atomic Energy Commission.⁷¹ Some scientific institutions have COI policies designed to discourage research that they view as more likely to be biased, and many have COI-disclosure policies designed to encourage scientists to take sponsorship into account when evaluating each other’s work. An examination of an investigation’s funding source may help judges be alert to potential biases and assess their likely direction.

Misconduct

Beyond bias, scientific misconduct can occur when a scientist doesn’t fairly evaluate other scientists’ work, doesn’t honestly report methods and results, doesn’t fairly assign credit, or doesn’t work within the ethical guidelines of the community.⁷² These deceitful practices are penalized and discouraged in various ways, including with career curtailment, funding bans, fines, and potentially prison time. Scientific misconduct, as well as honest mistakes, can lead to the acceptance of inaccurate hypotheses. Such errors will be corrected as ideas are reexamined in different contexts, including in the course of a trial. For example, in *Schuene-man v. Arena Pharms., Inc.*,⁷³ the court found that a pharmaceutical company had engaged in misrepresentation when it claimed that animal studies showed its

70. Lesné et al., 2006, *supra* note 57.

71. Molina & Rowland 1974, *supra* note 7.

72. For more, see Charles Gross, *Scientific Misconduct*, 67 Ann. Rev. of Psych. 693–711 (2016), <https://doi.org/10.1146/annurev-psych-122414-033437>.

73. 840 F.3d 698 (9th Cir. 2016).

drug to be noncarcinogenic, when in fact the company was conducting a study that linked the drug to cancer in lab rats, a case in which scientific results were not honestly reported. Another well-known example of scientific misconduct on trial involves Takata Corporation submitting fraudulent data on the performance of its airbag inflators to automakers. Takata pled guilty to wire fraud in 2017, nearly 20 years after its internal research had shown that its inflators did not perform to required specifications.⁷⁴

Scientific misconduct can be difficult to detect; there are no simple red flags that a judge could identify that would indicate misconduct.⁷⁵ For example, misconduct may be reported by someone with firsthand knowledge of the environment in the research group, it may be detected by applying specialized tools for image analysis to publications, or it may be detected by subject matter experts who notice a discrepancy or anomaly in reported data or cannot replicate a result. An allegation of misconduct will generally lead to an extensive investigation and may take many years to resolve and correct. A judge or jury might justifiably be concerned that a scientist who has engaged in documented misconduct may be an unreliable source of expert testimony.

Correction and Retraction

After peer review and publication, a scientific paper may be corrected if a small error that does not invalidate the main findings of the paper is caught—for example, an incorrect row in a data table, missing text, or a missing citation. In such cases, a publisher will generally update the version of record available online and associate a correction notice with the document so that readers have a record of what changes have been made and why. Guidelines and policies for error corrections may vary across journals and are generally intended to maintain the validity of the scientific record, though it has been argued that journals and editors may not have made sufficient efforts in this regard.⁷⁶ The fact that a paper has had a correction made should not be taken to indicate that the study is of poor quality. Instead, it suggests that the article has received additional scrutiny after publication and that the authors and publishers have made public efforts to address concerns

74. *In re Takata Airbag Prods. Liab. Litig.*, 193 F. Supp. 3d 1324 (S.D. Fla. 2016).

75. The Office of Research Integrity of the U.S. Department of Health and Human Services has identified five red flags of research misconduct, but all are observations that would only be made by other scientists working in the same field or by other workers with a professional relationship to the scientist suspected of misconduct. Office of Research Integrity, *Possible Red Flags of Research Misconduct* (2016), <https://perma.cc/JA9J-XKTE>.

76. Lonni Besançon et al., *Correction of Scientific Literature: Too Little, Too Late!*, 20 PLoS Biology e3001572 (2022), <https://doi.org/10.1371/journal.pbio.3001572>; Ambar Castillo, *Mistakes Happen in Research Papers. But Corrections Often Don't*, STAT (Jan. 10, 2023), <https://perma.cc/TQX7-3YXW>.

raised by that scrutiny, which can be viewed as a strength. Corrections in scientific papers are rare and have remained relatively constant since the 1970s.⁷⁷

On the other hand, if evidence arises that the entire basis of a published study is invalid, the paper may be retracted. Articles may be retracted because of a significant flaw that changes the conclusions of the paper (e.g., faulty equipment used to collect data or an error in code used to analyze the data) or misconduct of many varieties—plagiarism, fraudulent data, ethics violations, or manipulations of the peer-review process.⁷⁸ Theoretically, retraction means removal from the scientific record. In practice, however, the article is likely to remain available in libraries and online tagged with a notice of retraction, meant to signal to readers that the journal no longer backs the content of that article. One reason these papers remain available is because of the cumulative nature of science: Researchers may need to review retracted papers to see if the errors within them have propagated into other work. Because such articles remain accessible, retracted research may continue to garner citations improperly.⁷⁹ While the number and frequency of retractions from scientific journals has been on the rise, multiple analyses suggest that this is largely due to improved oversight and new tools to detect plagiarism and image manipulation.⁸⁰ Studies are retracted because they do not emerge from sound scientific methodologies, and thus any science that has been retracted should be disallowed in expert witness testimony.

Within science, retractions may be viewed as a blot on one's record of scholarship, and the scientist behind a retracted study may experience stigma from the scientific community.⁸¹ Should the same be true in the courtroom? Retraction is not the same as scientific misconduct, in which there is an intent to deceive. While scientific misconduct is the most frequent cause of retraction, at least in the life sciences, it is by no means the only cause.⁸² The circumstances that can lead to

77. Daniele Fanelli, *Why Growing Retractions Are (Mostly) a Good Sign*, 10 PLoS Med. e1001563 (2013), <https://doi.org/10.1371/journal.pmed.1001563> [hereinafter Fanelli 2013].

78. William Bülow, et al., *Why Unethical Papers Should be Retracted*, 47 J. Med. Ethics (2021), <https://doi.org/10.1136/medethics-2020-106140>.

79. Tzu-Kun Hsiao & Jodi Schneider, *Continued Use of Retracted Papers: Temporal Trends in Citations and (Lack of) Awareness of Retractions Shown in Citation Contexts in Biomedicine*, 2 Quantitative Sci. Stud. 1144–69 (2022), https://doi.org/10.1162/qss_a_00155.

80. Jeffrey Brainard & Jia You, *What a Massive Database of Retracted Papers Reveals About Science Publishing's 'Death Penalty'*, Science (Oct. 25, 2018), <https://doi.org/10.1126/science.aav8384>; Fanelli 2013, *supra* note 77; and Ferric C. Fang, R. Grant Steen & Arturo Casadevall, *Misconduct Accounts for the Majority of Retracted Scientific Publications*, PNAS Early Edition (2012) [hereinafter Fang et al. 2012].

81. This stigma may be detrimental to scientific knowledge building because it discourages the reporting of honest errors; for example, see Jaime A. Teixeira da Silva & Aceil Al-Khatib, *Ending the Retraction Stigma: Encouraging the Reporting of Errors in the Biomedical Record*, 17 Rsch. Ethics 251–59 (2021), <https://doi.org/10.1177/1747016118802970>.

82. Fang et al. 2012, *supra* note 80.

retraction include honest mistakes (which may initially be known only to the researcher who made them), reported out of respect for the knowledge-building function of science. In such cases, acknowledgment of the error might reasonably be perceived as a sign of integrity and commitment to developing accurate scientific explanations. For example, in 2006, pharmaceutical scientist Geoffrey Chang and colleagues retracted five papers that relied upon a flawed computer program, a mistake they owned up to and corrected.⁸³ Chang then continued his research, making valuable contributions to our understanding of microbial resistance, malaria, and crop development, with no indications that his new findings should be discounted because of a previous error. However, a pattern of retracted research stemming from a single researcher or group and tied to allegations of misconduct or bad science *can* suggest that those practitioners are not adhering to the norms of the scientific community, making their unretracted research suspect as well.

Retraction does not occur if the hypothesis a study supports is simply found to be incorrect by subsequent research. The methods and data of that study may still hold valuable insights for researchers, who can go back to it to try to understand why a different conclusion was reached in that case. This is a normal part of the self-correcting process of science.

Scientific Expertise

Judges may need to evaluate the qualifications and expertise of individual scientists to help determine whether to admit scientific testimony. On what basis can this judgment be made? Scientific experts are respected and active participants in their scientific communities, usually contributing to knowledge development as researchers and certainly scrutinizing the work of others as consumers of the research literature. Just as there is no single trait that separates science from nonscience, there is no single necessary and sufficient condition that would qualify a scientist as an expert on a particular topic in the eyes of other scientists. Instead, judges may evaluate a cluster of traits that a scientist is likely to attain in the process of contributing to a field and participating in the scientific community. These include:

- holding an advanced degree, usually a Ph.D., in a field closely related to the area of purported expertise,
- holding a position in which reviewing the latest research on the topic is required (e.g., having a specialized medical practice, an editorial position at a scientific journal, or a research career),

83. Greg Miller, *A Scientist's Nightmare: Software Problem Leads to Five Retractions*, 314 *Science* 1856–57 (2006), <https://doi.org/10.1126/science.314.5807.1856>.

How Science Works

- holding leadership positions in relevant scientific societies or institutions, or having other markers of professional esteem by the community,
- having published peer-reviewed research on the topic, and
- developing a track record, often measured in citation metrics, of publications that are well respected by colleagues and contribute to knowledge development.

Citation metrics, or citation impacts, are statistics that indicate how often a particular academic work, author, or journal has been cited. These metrics range from simple tallies to more complex statistics, such as the h-index, which incorporates information about the distribution of citations across an author's publication record.

Lacking a few of the traits in the list above does not disqualify an individual as a scientific expert, but lacking all or most of them should lead one to question whether the proposed expert is indeed an active, expert member of the relevant scientific community. Sherwood Rowland and Mario Molina had accumulated all of these markers of expertise by the end of their careers, of course, but Molina was an early career scientist at the time the CFC/ozone hypothesis was published, with only a few publications and a Ph.D. Had he been proposed as an expert witness on the ozone layer at that time, a judge might reasonably have questioned his expertise.

It is also worth evaluating the closeness of a scientist's disciplinary expertise to the scientific topic on which expert testimony will be delivered. Doctors and honorifics abound, and promoters of pseudoscientific ideas or hypotheses with little evidence supporting them may leverage achievements in other areas to create the impression of relevant expertise. For example, Richard Scorer, a well-known scientist, was promoted as an expert on the CFC/ozone hypothesis by an industry-backed group aiming to discredit the hypothesis.⁸⁴ Scorer had a Ph.D., had received scientific honors, and had conducted research on air pollution; however, he had no expertise in atmospheric chemistry or the ozone layer and had published no peer-reviewed articles on this topic. Scorer was, in fact, an expert in some related disciplines, but lacked the specialized knowledge needed to fairly evaluate the CFC/ozone hypothesis in light of the available evidence and research. The problem of scientists with legitimate expertise in one field weighing in on a scientific question outside their area of expertise is a pernicious one that has affected public acceptance of science and policy on issues such as climate change and tobacco exposure.⁸⁵ This scenario may play out in courts as judges identify and disallow testimony from scientists whose expertise is legitimate but not relevant to the specific scientific issue at hand. Many of the later reference

84. Dotto & Schiff 1978, *supra* note 69; Parson 2003, *supra* note 69.

85. Oreskes & Conway 2010, *supra* note 2.

guides in this manual provide guidance on qualifications for experts in particular disciplines.

Testing Hypotheses

At the most basic level, scientists test their ideas by figuring out what predictions are generated by a hypothesis and comparing those to evidence. Hypotheses are supported when the evidence matches predictions and are contradicted when they do not match.⁸⁶ Observations are often made with the help of sophisticated tools and may be analyzed with statistical techniques to discern patterns. Other observations are entirely qualitative and nonnumeric, taking the form of case studies or interviews, for example. Here, we review some overarching approaches to testing hypotheses that cut across all disciplines and address common concerns relating to the interpretation of data from such tests.

Scientific Methodologies and Data

The methodologies used in different scientific disciplines are so varied that it might seem a challenge to identify any fundamental similarities. How can it be that investigations as diverse as measuring ozone levels over Antarctica, exposing mice to cigarette smoke tars, and surveying consumers about their perception of the word *thermos* all serve the same fundamental purpose of testing hypotheses? In fact, scientific methods used to test hypotheses can be grouped in a few common ways, and understanding these can help in evaluating the strengths of individual studies. These attributes include:

- whether or not the research involves an intentional manipulation of condition (i.e., Is the research experimental or nonexperimental in nature?)
- what types of data the investigation produces (i.e., Did the research generate quantitative data, qualitative data, or both?)
- whether or not the hypotheses being examined are reflected in a mathematical model.

There are no right, wrong, better, or worse answers to the above questions. Instead, these attributes help determine the types of hypotheses and approaches best suited to gather evidence, as described below. Note that these characteristics vary relatively independently of each other and of scientific discipline: Experimental or nonexperimental methods can be used to collect quantitative or

86. Michael Strevens, *The Knowledge Machine: How Irrationality Created Modern Science* (2020).

qualitative data, with or without incorporating modeling, and these methodologies are applied broadly across the many topics that science investigates. Later reference guides in this manual explain how these methodological approaches are used in various disciplines.

Experimental and Nonexperimental Methods

An experiment involves intentionally manipulating some factor in a system to learn how that affects the outcome.⁸⁷ A controlled experiment seeks to keep variables other than the experimental manipulation the same in order to isolate the cause of any change. A control group, on the other hand, is a comparison condition or group that does not receive the experimental manipulation to increase confidence that the change observed in the experimental group is caused by the manipulation and not some other factor. For example, experiments examining whether the drug Bendectin causes birth defects, the basis of the *Daubert* litigation, compared experimental groups of animals given Bendectin during pregnancy to control groups of animals not given the drug.⁸⁸ Experiments can be performed in a lab or in a natural setting. For example, economic decision making may be investigated by bringing participants into a psychology lab and observing them playing a game, or by doing a field experiment at a market or swap meet.⁸⁹

A randomized controlled trial is an experimental approach that randomly assigns participants to the experimental or control group to better control variables across the groups. It is most commonly used in medical trials, but can be applied more broadly as well, for example to examine the effectiveness of a kindergarten curriculum at reducing bullying.⁹⁰ Randomized controlled trials can be very convincing and are often perceived as the gold standard in medicine.⁹¹ However, they are just one line of evidence among many.

87. Nancy Cartwright, *Nature's Capacities and Their Measurement* (1989).

88. For an example, see Rochelle W. Tyl et al., *Developmental Toxicity Evaluation of Bendectin in CD Rats*, 37 *Teratology* 539 (1988), <https://doi.org/10.1002/tera.1420370603>. For an overview, see Joseph Sanders, *From Science to Evidence: The Testimony on Causation in the Bendectin Cases*, 46 *Stanford L. Rev.* 1–86 (1993), <https://doi.org/10.2307/1229235>.

89. For example, see Werner Güth & Oliver Kirchkamp, *Will You Accept Without Knowing What? The Yes-No Game in the Newspaper and in the Lab*, 15 *Experimental Econ.* 656–66 (2012), <https://doi.org/10.1007/s10683-012-9319-7>; and Steven D. Levitt & John A. List, *Field Experiments in Economics: The Past, the Present, and the Future*, 53 *Eur. Econ. Rev.* 1–18 (2009), <https://doi.org/10.1016/j.eurocorev.2008.12.001>.

90. For example, see Adele Diamond et al., *Randomized Control Trial of Tools of the Mind: Marked Benefits to Kindergarten Children and Their Teachers*, 14 *PLoS ONE* e0222447 (2019), <https://doi.org/10.1371/journal.pone.0222447>.

91. For example, see *In re Bextra & Celebrex Mktg. Sales Practices & Prod. Liab. Litig.*, 524 F. Supp. 2d 1166 (N.D. Cal. 2007).

Controlled experiments are often misperceived as the pinnacle of scientific evidence, with all other forms of testing ranked as substandard in some way. Experiments *are* a useful way to collect evidence, but many important questions cannot be addressed through experiment. Some aspects of the natural world aren't manipulable (e.g., distant stars), and hence can't be studied with direct experiments. In other cases, it would be unethical to perform a direct experiment on the study system—for example, randomly assigning some pregnant people to receive Bendectin to assess its safety or intentionally introducing a pollutant to a sensitive ecosystem to determine its effect.

Nonexperimental methods, which involve no intentional manipulation of condition, are often critical in gathering evidence to evaluate important hypotheses. These methods still allow us to generate predictions from a hypothesis and make observations to test those predictions. For example, we can't experiment on stars to test hypotheses about which elements occur within them, but we *can* test those ideas by figuring out what wavelengths of light the presence of different elements would generate and monitoring the stars' emission spectra.⁹²

Nonexperimental research describes or documents a phenomenon. We might imagine a biologist sketching a bacterium viewed through a microscope, but of course, complex machinery, custom-designed instruments, and detailed measurements may also be involved. Analyses to test drinking water for contaminants, detect radiation coming from deep space, or measure consumer perception of a particular brand (e.g., surveys to support trademark litigation) fall into this category. Often nonexperimental studies involve making some sort of comparison among observations to detect patterns and associations. For example, a key study supporting the CFC/ozone link documented ozone measurements over Antarctica and compared results obtained between 1957 and 1984, showing a pattern of dramatic decline.⁹³ This key piece of convincing evidence was not obtained through an experiment, but through observation and comparison over time. Similarly, many nonexperimental epidemiologic studies informed the *Daubert* litigation, some that compared birth outcomes for mothers who had taken Bendectin to those who did not (cohort studies) and some that compared Bendectin exposure between those with a birth defect and those without (case-control studies).

Some effects we might wish to study experimentally cannot be examined in that context because of ethical or logistical considerations. However, scientists can leverage situations in which such changes occur naturally, without an intentional manipulation on the part of the researcher, to gather relevant data in a nonexperimental setting. For example, studying the connection between years

92. For example, see Noriyuko Matsunaga et al., *Identification of Absorption Lines of Heavy Metals in the Wavelength Range 0.97-1.32 μm* , 246 *Astrophysical J. Supp. Series* 10 (2020), <https://doi.org/10.3847/1538-4365/ab5c25>.

93. Farman et al. 1985, *supra* note 8.

of parental schooling and children's well-being presents nearly insurmountable experimental challenges. However, researchers have taken advantage of variation in the rollout of government educational programs and policies to find comparison groups that mimic the desired experimental setup, contrasting outcomes from children whose parents received extra years of schooling with those from closely matched communities that received the program later.⁹⁴ More recently, in another comparative, nonexperimental study, researchers used data from school districts that lifted mask mandates at different times to examine the on-the-ground effect of masking on transmission of SARS-CoV-2.⁹⁵

Many explanations can be tested through both experimental and non-experimental methods. Each offers advantages. A laboratory experiment may allow a researcher to control many factors and make a strong case for causality; however, such experiments may not be able to closely mimic the complex environment outside the laboratory and so are vulnerable to the argument that the causal mechanism at work in the experimental setup is not important where it matters most—in the messy real-world conditions found in places like hospitals, schools, and natural ecosystems. On the other hand, a nonexperimental, comparative study of a field setting may be clearly grounded in those real-world conditions but leave open questions about confounding causal factors. For example, medical studies using nonhuman animal models may provide convincing evidence of the carcinogenic effects of a substance in the model organism but not necessarily people, while epidemiologic evidence of cancer rates in people are clearly relevant to human cancers but may be consistent with multiple hypotheses as to the cause of the cancer (see the reference guides on toxicology and epidemiology, in this manual, for further considerations). The nonexperimental masking study referenced above addressed some complaints about lab-based mask research not reflecting real-world variation in the types of masks worn in schools and the consistency with which they are worn, but the study was nevertheless criticized by opponents for not being a randomized controlled trial.⁹⁶

Whenever possible, triangulating across a body of evidence based on multiple methodologies is the best approach to evaluating the scientific support for a hypothesis. The scientific community does not fully exclude consideration of a particular type of evidence relevant to a hypothesis unless it is clearly compromised methodologically or there are other a priori reasons for thinking it

94. Shin-Y Chou et al., *Parental Education and Child Health: Evidence from a Natural Experiment in Taiwan*, 2 Am. Econ. J.: Applied Econ. 63–91 (2010), <https://doi.org/10.1257/app.2.1.33>. For more on the complex statistical models that can be used to analyze data from studies such as this one, see David H. Kaye and Hal S. Stern, *Reference Guide on Statistics and Research Methods*, in this manual.

95. Tori L. Cowger et al., *Lifting Universal Masking in Schools—Covid-19 Incidence among Students and Staff*, 387 New Eng. J. Med. 1935–46 (2022), <https://doi.org/10.1056/NEJMoa2211029>.

96. Roni Caryn Rabin, *Masks Cut Covid Spread in Schools, Study Finds*, N.Y. Times (Nov. 10, 2022), <https://perma.cc/JVJ5-LUTX>.

irrelevant. This same issue, assessing the value of the many lines of evidence relevant to a particular question (e.g., if and how nonexperimental epidemiologic studies and experimental animal studies should inform deliberation in toxic tort litigation), has presented challenges for the courts and led to decisions about admissibility that have generated controversy.⁹⁷ In science, the available evidence (some of which may come from other research programs not designed to test the hypothesis under consideration) is evaluated as a body, along with the strengths, weaknesses, and caveats relating to each type of data, an approach which, some scholars have argued, the judiciary has not always followed.⁹⁸ When a particular experimental or nonexperimental methodology has been applied to a question multiple times, special analytic techniques like meta-analyses may be used to interpret this body of evidence (see section titled “Systematic Reviews and Meta-analyses” below).

Within scientific methodologies, blinding generally refers to concealment of comparison group membership so that this knowledge does not bias the study outcome. In a double-blind clinical trial, neither the researcher nor the participants know which group is receiving a placebo and which is receiving the experimental intervention until after data collection is complete. Triple blinding refers to additionally concealing this information from the data analysts. In non-comparative studies, blinding may refer to concealment of the purpose or sponsor of the research from participants and those implementing the protocol (see the *Reference Guide on Survey Research*, in this manual). Blinding can be helpful in many investigations, experimental or not, in which human perception might tip the outcome in a particular direction—for example, in handwriting analysis and in proficiency testing to determine a lab’s ability to perform a particular

97. For an overview, see Margaret A. Berger, *What Has a Decade of Daubert Wrought?*, 95 Supp. 1 Am. J. Pub. Health s59–s65 (2005) [hereinafter Berger 2005]. For a small sample of discussion relevant to this issue, see Gary Edmond & David Mercer, *Litigation Life: Law-Science Knowledge Construction in (Bendectin) Mass Toxic Tort Litigation*, 30 Soc. Stud. Sci. 265–316 (2000); Victoria Sutton & Brie DeBusk Sherwin, *Toxicological Animal Studies Disparate Treatment as Scientific Evidence*, 2 J. Animal & Env’t L. (2011), <https://perma.cc/V2U3-SXKV>; and Raymond Richard Neutra, Carl F. Cranor, & David Gee, *The Use and Misuse of Bradford Hill in U.S. Tort Law*, 58 Jurimetrics 127–62 (2018) [hereinafter Neutra et al. 2018]. For argument on the opposing side, see, for example, Dije Ndreu Comment, *Keeping Bad Science Out of the Courtroom: Why Post-Daubert Courts Are Correct in Excluding Opinions Based on Animal Studies from Birth-Defects Cases*, 36 Golden Gate U. L. Rev. 459 (2006); and Amanda Hungerford Note, *Back to Basics: Courts’ Treatment of Agency Animal Studies after Daubert*, 110 Colum. L. Rev. 70–113 (2010).

98. Some scholars have raised concerns that the courts have on occasion unfairly dismissed numerous individual lines of evidence as being flawed or insufficiently conclusive and concluded that evidence is lacking, when in fact the *body* of evidence, taken as a whole, points to a clear conclusion. For more, see discussion of *Milward v. Acuity Specialty Products Group, Inc.*; see also Liesa L. Richter & Daniel J. Capra, *The Admissibility of Expert Testimony*, in this manual; Berger 2005, *supra* note 97; and Steve C. Gold, *A Fitting Vision of Science for the Courtroom*, 3 Wake Forest J.L. & Pol’y 1 (2013).

technique (see the *Reference Guide on Forensic Feature Comparison Evidence*, in this manual).

Qualitative Data, Quantitative Data, and Mixed Methods

Quantitative data are numeric; qualitative data are nonnumeric, taking the form of, for example, written descriptions, audio recordings, and video footage. Both are essential forms of evidence within science, just as they are in a trial. Qualitative data are often misperceived as soft, less useful than quantitative data, and the domain of the social sciences. This view overlooks many applications of qualitative data within the natural sciences—for example, to visually identify biological species, minerals, and strata. Neither are quantitative data absent from the social sciences, as illustrated by numerous examples in the *Reference Guide on Survey Research*, in this manual. Furthermore, nuanced qualitative data from social science, psychology, and behavioral science are particularly valuable for providing context and meaning for quantitative results, as well as for suggesting hypotheses for further investigation. For some analyses, qualitative data may be validly aggregated and converted to quantitative data through a coding scheme. Mixed-methods research uses a combination of qualitative and quantitative techniques within a single investigation to detect patterns and form and test hypotheses.

When research is quantitative in nature, scientists often use common numeric standards to evaluate the outcomes of their tests:

- *Evaluating hypotheses:* In Bayesian approaches, scientists attempt to calculate the probability that a hypothesis is true given the data available. Such approaches begin by choosing a *prior probability distribution* (“priors”) over mutually exclusive hypotheses. These priors are based on previous research, background theories, or even statistical indifference between possibilities. Then as observations are made or experiments conducted, a new probability distribution called the *posterior* distribution is calculated using Bayes’s theorem. Over repeated observations or experiments, the values of the posterior distribution will converge to the true values. Thus, Bayesian approaches allow scientists to calculate the probability that the hypothesis is true given the evidence they have observed. In contrast, in the commonly used *p*-value approach, scientists compare a test hypothesis (e.g., that drug X is effective) to a null (e.g., that there is no difference in cure rates between those who took drug X and those who took a placebo). Scientists then calculate the probability that the null hypothesis could be true even with the observed difference between conditions (e.g., the cure rate of patients taking drug X compared to that of those taking a placebo). For more details, see the *Reference Guide on Statistics and Research Methods*, in this manual. In these cases, scientists

often use a probability of 0.05 as the threshold of significance, indicating that the null hypothesis would generate a result as or more extreme than the observed data only 5% of the time, providing support to the non-null hypothesis. For example, in Zoloft product liability litigation, the court excluded expert testimony (arguing that Zoloft caused birth defects in the children of mothers taking the drug) because the testimony was based on trends—patterns in data that do not reach the level of statistical significance—leaving a strong possibility that any association between Zoloft and birth defects was due to chance alone (the null hypothesis).⁹⁹ When considering testimony and evidence that relies on p -values, it is important to keep in mind a few caveats:

- (1) A p -value lower than 0.05 does not *prove* that a null hypothesis is false. It is strong evidence, but there is a small chance that the difference observed could be the result of chance alone.
- (2) Using a low p -value (e.g., 0.05) as a criterion for significance sets a high bar for rejecting the null hypothesis, minimizing the chance of getting a false positive—as a hypothetical example, minimizing the chance that we conclude that Zoloft use in pregnancy leads to more birth defects when, in fact, there is no relationship between Zoloft and birth defects. When the p -value is higher than the cutoff criterion and the null is not rejected (and note that the lower our cutoff criterion, the more likely this is to happen), it might be tempting to conclude that the null hypothesis must be true (e.g., that there is no relationship between Zoloft and birth defects). But this is erroneous. Having not enough evidence to establish a relationship is not the same as having strong evidence that there is no relationship. We should also be concerned about false negatives—that is, concluding that Zoloft use in pregnancy has no impact on birth defects in the hypothetical case that Zoloft use *does* in fact lead to more birth defects—which could have drastic implications for toxic tort litigation. The likelihood of false negatives can be assessed through an examination of statistical power (see paragraph (3) below).
- (3) Depending on its design and the underlying system, a study may have limited ability to reject a null hypothesis (i.e., detect a difference) at the significance cutoff criterion. Power refers to a test's ability to reject a hypothesis that is indeed false. A study with high statistical power minimizes the chance of false negatives. Statistical power increases with the sample size of the study and with the effect size one wishes to detect (see Evaluating impact bulleted paragraph below). Well-designed studies have sufficient power to detect the differences of interest, but it may not be apparent when a test lacks power (see the *Reference Guide*

99. *In re Zoloft (Sertraline Hydrochloride) Prods. Liab. Litig.*, 26 F. Supp. 3d 449 (E.D. Pa. 2014).

on *Statistics and Research Methods*, in this manual, for more details), and it is possible to leverage underpowered studies to argue for no causal effect when one does in fact exist.¹⁰⁰

- (4) When several separate studies find nonsignificant trends, meta-analyses (see section titled “Systematic Reviews and Meta-analyses” below) may be used to examine what the data taken as a whole say, as any individual test may have limited statistical power for the reasons described above.
 - (5) The 5% cutoff value for significance is a convention, not a number that emerges out of nature or statistical theory, and scientists have their own debates about how p -values should be used.¹⁰¹ There is nothing special about 0.05 other than the fact that it is an a priori criterion; clearly 0.04 or 0.06 are not very different. As described in the section titled “Assessment of Evidence” below, scientists may take into account the outcome of accepting or rejecting a particular hypothesis and adjust the design of their tests, or their threshold for action or acceptance, to guard against harm.
- *Evaluating impact:* While p -values and Bayesian approaches attempt to answer the question “Is there a relationship between variables?” effect sizes answer the question “How strong is this relationship?” P -values can generally be used to interpret the significance of a relationship in the same way across studies, but effect size can be communicated using many different statistics. Effect sizes are an important consideration at trial because they provide an indication of how much of an outcome can be attributed to a particular causal agent. For example, congenital cardiovascular defects in neonates can be caused by a variety of factors and occur relatively frequently even when mothers did not take an antidepressant during pregnancy. In the Zolof case mentioned above, one of the studies cited by the memorandum opinion found that a different antidepressant, fluoxetine, had an adjusted odds ratio of 4.47, meaning that mothers who took fluoxetine during early pregnancy were estimated to be 4.47 times more likely to have a baby with a congenital cardiovascular anomaly than mothers who did not take the antidepressant during early pregnancy.¹⁰² Causal relationships with a weak effect are difficult to detect without very large

100. See Neutra et al. 2018, *supra* note 97, for discussion of the judiciary’s awareness of power when evaluating statistical significance.

101. For example, see Daniel J. Benjamin et al., *Redefine Statistical Significance*, 2 *Nature Hum. Behav.* 6–10 (2018), <https://doi.org/10.1038/s41562-017-0189-z>; John P.A. Ioannidis, *The Importance of Predefined Rules and Prespecified Statistical Analyses: Do Not Abandon Significance*, 321 *JAMA* 2067–68 (2019), <https://doi.org/10.1001/jama.2019.4582>; Ronald L. Wasserstein, Allen L. Schirm & Nicole A. Lazar, *Moving to a World Beyond “ $p < 0.05$ ”*, 73 *suppl Am. Statistician* 1–19 (2019), <https://doi.org/10.1080/00031305.2019.1583913>.

102. Orna Diav-Citrin et al., *Paroxetine and Fluoxetine in Pregnancy: A Prospective Multicentre, Controlled, Observational Study*, 66 *Brit. J. Clinical Pharmacology*, 695–705 (2008), <https://doi.org/10.1111/j.1365-2125.2008.03261.x>.

sample sizes. In the Zolof litigation, the plaintiffs' expert's testimony would have argued that multiple, nonsignificant associations between Zolof use and birth defects indicated a causal relationship. The testimony was excluded because these results were consistent with a weak causal relationship (a small effect size), one that is "so weak that one cannot conclude that the risk is greater than that seen in the general population."¹⁰³

- *Evaluating estimates:* In science (and in contrast to their lay meanings), the terms *uncertainty* and *error* refer to the variability of a set of data that is intended to estimate a single number. Uncertainty and error are generally expressed as a range, within which we are confident that, if the study were repeated, the new result would fall. Scientists often use a 95% confidence interval for this purpose. For example, the researchers who first documented the ozone hole compiled daily measurements of total ozone over Halley Bay, Antarctica, taken over more than two decades. The measurements varied substantially from one day to the next. To communicate the overall pattern, the researchers estimated the maximum error bounds on their measurements and showed that the trend of decreasing ozone levels over time was much larger than any potential errors in the measurements. No matter where within the error range the true ozone values fell, the pattern they'd identified would still be exhibited.

Note that the 95% and 5% cutoffs are somewhat arbitrary, and a higher degree of confidence might be required if more certainty were desired—for example if an impactful policy decision depended on the conclusion. Similar, cross-disciplinary, standardized cutoff values for effect sizes do not exist because of the many different statistics used to communicate effect size and because interpreting these statistics is highly dependent on the system under consideration and for what purposes one needs to interpret associations. For more on these statistics, see the *Reference Guide on Statistics and Research Methods*, in this manual.

Importantly, a low *p*-value, large effect size, or narrow confidence interval found in a particular study should not be taken to indicate scientific consensus. These statistical measures help communicate the results of an isolated investigation, and scientific consensus is built over the course of many investigations and analyses (see section titled "Achieving Scientific Consensus" below).

Modeling

The term *model* is used in many different ways in science. However, as a research method of modern science, modeling almost always refers to developing and testing

103. Memorandum Opinion 9, *In Re Zolof Products Liability Litigation*, No. 2:12-MD-2342 at 9 (E.D. Pa. 2014), ECF No. 979.

a mathematical model. These models can be relatively simple and thus analytically tractable; however, other models are so complex that they are built as a piece of computer code. One can think of a model as a surrogate, a simplified representation of a real-world system.¹⁰⁴ That system could be any part of the natural world, from how gases move in the atmosphere to how social media influencers arise.¹⁰⁵

Models generally accept multiple different input values, or parameters, and then generate a set of predictions, which are usually, though not always, quantitative in nature. For example, models of Earth's atmosphere and climate¹⁰⁶ accept input about the shape of terrain, vegetation cover, ice cover, atmospheric makeup, and many other factors. The models integrate mathematical rules that simulate how gases circulate over the surface of earth, how sunlight is reflected from various surfaces, how water moves from land surfaces into the atmosphere, and more. And they generate predictions about future states of the Earth system: temperature ranges, precipitation levels, frequency of extreme weather events, etc. Essentially, a model presents an argument: If these input values represent the starting point of the system, and if the system works according to the rules implemented in the model's code or equations, then we'd expect the system to have this predicted state in the future.

Models can and must be tested against evidence, just like hypotheses. Their predictions can be compared to observations to see if the model seems to be a fair representation of the system. If a model's predictions do not match key elements of our observations or evidence, we must reevaluate our input parameters and/or the sufficiency and accuracy of the model.¹⁰⁷ The atmospheric models that were used to test Rowland and Molina's hypothesis incorporated many different units of scientific knowledge, which were themselves supported by other lines of evidence. While the models were not perfect representations of the Earth system, they were close enough to generate many predictions that held true of the atmosphere in general, giving scientists confidence that the models were useful approximations of actual mechanisms. And when Rowland and Molina's hypothesis was incorporated into those models, they generated two key predictions that were borne out in observations: that CFCs should be abundant in the lower atmosphere but rare in the upper atmosphere, a prediction that was immediately verified with evidence from sensors placed on aircraft and balloons,

104. Michael Weisberg, *Simulation and Similarity: Using Models to Understand the World* (2013) [hereinafter Weisberg 2013].

105. For example, see Penelope Maher et al., *Model Hierarchies for Understanding Atmospheric Circulation*, 57 *Rev. Geophysics* 250–80 (2019), <https://doi.org/10.1029/2018RG000607>; and Nicolò Pagan et al., *A Meritocratic Network Formation Model for the Rise of Social Media Influencers*, 12 *Nature Comm'ns* 6865 (2021), <https://doi.org/10.1038/s41467-021-27089-8>.

106. For example, see June-Yi Lee et al., *Future Global Climate: Scenario-Based Projections and Near-Term Information*, in *Climate Change 2021: The Physical Science Basis* 553–672 (2021), <https://doi.org/10.1017/9781009157896.006>.

107. Weisberg 2013, *supra* note 104.

and that the ozone layer would become thinner over time, a prediction that was verified many years later.¹⁰⁸

If a model is supported and seems to be a good representation of the target system, we can use it to answer “what if” questions: What would have happened in 50 years if we had continued CFC production at 1974 rates? How would changing a social network’s recommendation system affect influencers’ ability to spread information? If a hazardous chemical is used in a continuous action air freshener, what dose would an individual confined mostly to the home inhale daily? If a business had continued to operate instead of being barred from doing so, what profits and value would it have accrued during that period of closure? Predictions from widely accepted models may be considered evidentiary and used to build other arguments—for example, about the cause of a disease or injury. The *Reference Guide on Exposure Science*, in this manual, explains the types of models that are used to estimate the exposures to and environmental fates of chemicals, and the *Reference Guide on the Estimation of Economic Damages*, in this manual, explains the types of financial models used to estimate damages.

Correlation and Causation

While the often-stated maxim that correlation does not imply causation is true, in fact, correlation is the only means that we have of establishing causation in science.¹⁰⁹ The reason we accept that smoking causes lung cancer is a series of very convincing correlations: between increases in population-level cigarette consumption and lung cancer rates; between exposure of lab animals to cigarette smoke tars and the development of tumors; between smoking and the presence of precancerous cells in the lungs; etc.¹¹⁰ Experiments are valuable for establishing causation because they allow us to rule out many confounding variables, but interpreting experimental evidence *still* involves looking for correlations between experimental manipulations and outcomes. When many correlations line up in different tests, all linked by a logical natural explanation, scientists are likely to accept a causal relationship.¹¹¹ In our ozone layer example, when correlations were shown between the presence of ice crystals and more rapid ozone-depleting

108. For example, see A. L. Schmeltekopf et al., *Measurements of Stratospheric CFC₁, CF₂Cl₂, and N₂O*, 2 *Geophysical Resch. Letters* 393–96 (1975), <https://doi.org/10.1029/GL002i009p00393>; and Farman et al. 1985, *supra* note 8.

109. Judea Pearl, *Causality* (2d ed. 2009).

110. Robert N. Proctor, *The History of the Discovery of the Cigarette-Lung Cancer Link: Evidentiary Traditions, Corporate Denial, Global Toll*, 21 *Tobacco Control* 87–91 (2011), <https://doi.org/10.1136/tobaccocontrol-2011-050338>.

111. For considerations in inferring causation in the field of epidemiology, see discussion of the Bradford-Hill criteria for causation in Steve C. Gold et al., *Reference Guide on*

chemical reactions in a lab environment, between the prevalence of icy polar clouds and the degree of ozone depletion in the stratosphere, and between the predictions of models that incorporated icy clouds and observations of the atmosphere, scientists accepted that polar stratospheric clouds were part of the cause of the extreme ozone loss they observed in Antarctica.¹¹² Correlations are the basis of essential evidence both in science and at trial.

Assumptions and Auxiliary Hypotheses

Assumptions are often perceived as a liability in logically reaching a sound conclusion, but in fact *all* scientific tests depend on making assumptions. Even testing a simple hypothesis about the presence of lead in drinking water involves making many assumptions about how the samples were collected, how the analytic equipment was calibrated, and of course, the accuracy of the stores of scientific knowledge that go into the field of optical emission spectroscopy, which is used to determine the elemental constituents of metals. One might wonder about trusting *any* scientific evidence if it all relies on assumptions. In science, assumptions are treated as auxiliary hypotheses, which can themselves be tested, a process that is part of the fabric of scientific knowledge building. By testing hypotheses and auxiliary hypotheses in different combinations, using different methods, science can triangulate between lines of evidence, homing in on the accuracy of a particular idea, independent of the assumptions that underlie its tests.¹¹³

For example, the sorts of atmospheric models that were used to examine the link between CFCs and ozone depletion are complex and themselves represent hundreds of auxiliary hypotheses—about how molecules move and diffuse through the atmosphere, about how solar radiation penetrates the atmosphere, about the strength of solar radiation throughout the day, and so forth—each of which had some justification and/or evidence supporting it.¹¹⁴ When Molina and Rowland discovered that chlorine nitrate was longer lived than previously reported and so could interact with chlorine atoms from CFCs, they published this information so that it could be added to atmospheric models.¹¹⁵ Several groups were involved in testing the Rowland–Molina hypothesis, and up to this

Epidemiology, in this manual. For an analysis of potential judicial interpretations of Bradford-Hill, see Neutra et al. 2018, *supra* note 97.

112. For example, see Susan Solomon, *Progress Towards a Quantitative Understanding of Antarctic Ozone Depletion*, 347 *Nature* 347–54 (1990), <https://doi.org/10.1038/347347a0>.

113. Carl G. Hempel, *Philosophy of Natural Science* (1966).

114. For an introduction to these models and examination of how they have developed over time, see Paul N. Edwards, *History of Climate Modeling*, 2 *WIREs Climate Change* 128–39 (2011), <https://doi.org/10.1002/wcc.95>.

115. Rowland et al. 1976, *supra* note 41.

point, their models' predictions had largely agreed. But once the researchers added the chlorine nitrate reactions to their models, different groups got conflicting results. Some models even predicted an increase in ozone. The discrepancy was traced to a particular auxiliary hypothesis. Those models predicting a net increase in ozone had one thing in common: They all used the approximation that the sun shines at its average intensity all the time, instead of varying throughout the day.¹¹⁶ Correcting this simplifying auxiliary hypothesis brought these models into agreement with the other models and with the actual atmospheric measurements of chlorine nitrate. Incorrect assumptions can be identified as such because of the iterative nature of the process of science.

Replication and the “Replication Crisis”

Scientists aim for their tests to be replicable—so that, for example, two studies examining consumers' perception of the “Thermos” brand name would yield concordant results if carried out in similar contexts.¹¹⁷ The goal of replicability comes with science's job of uncovering the mechanisms by which the natural world operates; these mechanisms are *consistently* at work, and so, if a study's finding cannot be replicated, it suggests that our understanding of the mechanism underlying the result or our testing methodologies are insufficient. In practice, many studies are not repeated step by step in an exact replication of the original, and those that are replicated exactly may not be published.¹¹⁸ Instead, hypotheses are often tested in slightly different contexts using varied methods, a process that helps establish the accuracy of the idea and further illuminates its scope and applicability. In our ozone layer example, scientists' use of multiple atmospheric models, some of which made simplifying hypotheses about the strength of solar radiation throughout the day and some of which did not, led to a better understanding of ozone depletion and of the sensitivity of the models, which are used for other purposes, like weather and climate predictions.

In recent years, concerns have arisen that many published scientific studies are unreplicable and that the hypotheses they purport to support may in fact be incorrect.¹¹⁹ This concern is often termed “the replication crisis,” and we will

116. F. S. Rowland, John E. Spencer, & Mario J. Molina, *Estimated Relative Abundance of Chlorine Nitrate among Stratospheric Chlorine Compounds*, 80 J. Phys. Chem. 2713–15 (1976), <https://doi.org/10.1021/j100565a020>.

117. In different disciplines and contexts, the terms *replicability*, *reproducibility*, and *repeatability* are used in inconsistent ways. Here, we use the term *replication* to mean a study carried out with the intent of mimicking another to the closest degree possible.

118. See NASEM, *supra* note 46, and Fiona Fidler & John Wilcox, *Reproducibility of Scientific Results*, in *The Stanford Encyclopedia of Philosophy* (Edward N. Zalta ed., 2021).

119. See NASEM, *supra* note 46, and more recent publications such as Timothy M. Errington et al., *Investigating the Replicability of Preclinical Cancer Biology*, 10 eLife e71601 (2021), <https://doi.org/10.7554/eLife.71601>.

use that terminology for consistency.¹²⁰ The identification of many unreplicable studies leads to valid questions about science's reward structure of promoting low quality and misleading practices. Individual scientists and scientific institutions have responded with proposals for incentivizing exact replication studies, incentivizing the reporting of negative or nonsignificant results, encouraging transparent reporting, diversifying peer review, and changing aspects of science's reward structure that seem to be to blame for the apparent proliferation of unreplicable studies.¹²¹ Concrete steps aimed at achieving these goals have been taken by, for example, expanding the use of platforms where researchers can preregister a planned investigation, describing hypotheses to be examined, methods used, and planned analyses. Preregistration may help reduce the likelihood that researchers will fail to report null results or mine their data for significant results, which will occur by chance alone 5% of the time with a p -value of 0.05, as described above. Preregistration is standard practice for medical clinical trials, but it is becoming more common, especially within psychology. However, it remains to be seen to what extent, if any, preregistration will improve replicability; research on this question is still in progress.¹²² Nevertheless, preregistration is just one of many initiatives that scientific institutions are pursuing to help address the replication crisis, reflecting their commitment to encouraging high-quality research.¹²³

While there is no question that changing science to promote higher-quality research is worthwhile and will make science more efficient at building accurate knowledge about the natural world, the fact that there is room for improvement in the process of science does not necessitate distrust of hypotheses that have gained widespread acceptance in the scientific community and about which consensus has been achieved. Indeed, some have argued that the problem of unreplicated results may not be any worse today than it has been over the history

120. However, as discussed herein, describing the situation as a "crisis" implies that it is a new situation and a catastrophic problem for scientific progress, which does not seem to be the case.

121. For examples of proposed changes, see Marcus R. Munafò et al., *A Manifesto for Reproducible Science*, 0021 Nature Hum. Behav. 1–9 (2017), <https://doi.org/10.1038/s41562-016-0021>; and Piers D. L. Howe & Amy Perfors, *An Argument for How (and Why) to Incentivize Replication*, 41 Behav. & Brain Scis. e135 (2018), <https://doi.org/10.1017/S0140525X18000705>, as well as NASEM, *supra* note 46.

122. For some research on this question, see A. Claesen, S. Gomes, F. Tuerlinckx, & W. Vanpaemel, *Comparing Dream to Reality: An Assessment of Adherence of the First Generation of Preregistered Studies*, R. Soc. Open Sci. 211037 (2021), <http://doi.org/10.1098/rsos.211037>, and Christopher Allen & David M. A. Mehler, *Open Science Challenges, Benefits and Tips in Early Career and Beyond*, 17 PLoS Biology e3000246 (2019), <https://doi.org/10.1371/journal.pbio.3000587>. For discussion of why preregistration might not work as planned, see Kai Kupferschmidt, *More and More Scientists Are Preregistering Their Studies. Should You?*, Science (2018), <https://doi.org/10.1126/science.aav4786>.

123. Max Korbmacher et al., *The Replication Crisis Has Led to Positive Structural, Procedural, and Community Changes*, 1 Comm'ns Psych. 1–13 (2023), <https://doi.org/10.1038/s44271-023-00003-2>.

of Western science and that, in any case, this problem is not impeding progress at building new, accurate knowledge about the natural world.¹²⁴

The replication crisis is a concern about individual investigations, not scientific consensus. Not all published ideas will be replicated, but neither are they all worth trying to replicate. Hypotheses that seem fruitful, explanatory, and potentially important are useful for understanding the world and so will be tested in different ways, by different people. In science, the fact that a particular study has not been precisely replicated does not mean it is invalid. Nor is a successful replication attempt necessarily an indication that the idea being tested is correct. Only ideas that have held up to repeated testing in different ways, having built up a body of supporting evidence, are likely to gain wide acceptance within science. However, judges are frequently asked to consider evidence that is far from achieving consensus. The replication crisis makes clear that peer review alone cannot ensure that the conclusions of published studies are actually correct, highlighting the responsibility judges bear in evaluating the validity of the methodologies that contributed to a particular piece of research.

Achieving Scientific Consensus

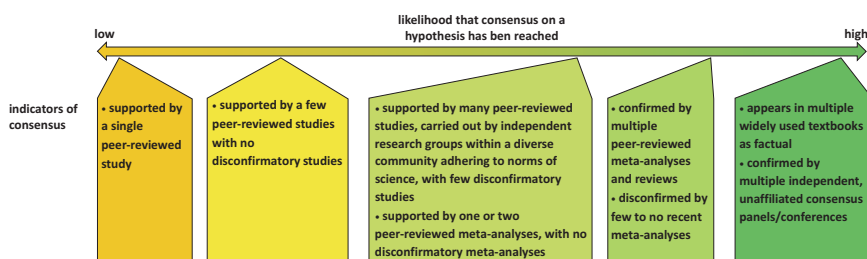
Much of the preceding text, as well as subsequent reference guides, will assist judges in evaluating the methods behind evidence relating to emerging or contentious science. Here, we examine the opposite end of the continuum of certainty: scientific consensus or widespread acceptance, the fifth factor that *Daubert* guides judges toward in their consideration of the reliability of expert testimony—though we emphasize that consensus is not a requirement for admissibility. While widespread acceptance provides a strong indicator of the reliability of scientifically acquired knowledge, there is no surefire way of assessing which hypotheses have reached this level of acceptance. No regular polls of scientists assess scientific consensus, although on some issues with significant societal importance and controversy, like climate change, such polls may be carried out.¹²⁵ Sometimes consensus conferences or panels are convened or consensus reports are written to assess the state of the field and/or to better communicate important conclusions to the public and policy makers. For example, several organizations, such as the

124. See Daniele Fanelli, *Is Science Really Facing a Reproducibility Crisis, and Do We Need It To?*, 115 PNAS 2628–31 (2018), <https://doi.org/10.1073/pnas.1708272114>; and Richard M. Shiffrin, Katy Börner, & Stephen M. Stigler, *Scientific Progress Despite Irreproducibility: A Seeming Paradox*, 115 PNAS 2632–39 (2018), <https://doi.org/10.1073/pnas.1711786114>.

125. For examples, see Peter T. Doran & Maggie Kendall Zimmerman, *Examining the Scientific Consensus on Climate Change*, 90 Eos 22–23 (2011), <https://doi.org/10.1029/2009EO030002>; Neil Stenhouse et al., *Meteorologists' Views About Global Warming: A Survey of American Meteorological Society Professional Members*, 95 Bull. Am. Meteorological Soc'y 1029–40 (2014), <https://doi.org/10.1175/BAMS-D-13-00091.1>.

Community Preventive Services Task Force and the Cochrane Collaboration, focus on producing reports that assess scientific consensus on medical questions, and the Intergovernmental Panel on Climate Change has, since 1990, produced reports assessing the scientific consensus on climate change. Similarly, the National Research Council and National Academies regularly produce consensus reports on a variety of scientific issues, including topics as diverse as cryptography, policing outcomes, and the health effects of electric and magnetic fields.¹²⁶ Returning to our ozone example, in 2003 the World Meteorological Association issued a report synthesizing scientific knowledge on ozone depletion to which 275 disciplinary experts contributed.¹²⁷ Such reports are valuable indicators of scientific consensus, but many scientific topics in question at a trial are unlikely to have such definitive documentation of acceptance to which to turn. Figure 3 summarizes some of the signals that may indicate scientific consensus has been reached, as well as signs that insufficient evidence has accumulated for the hypothesis to reach widespread acceptance. The link between CFCs and ozone depletion began at the left side of this diagram—as a single peer-reviewed study—and over the course of 15 years, accumulated so much evidence and withstood so many rounds of scrutiny that it has now achieved widespread acceptance and the highest level of certainty science has to offer, placing it at the far right side of the diagram.

Figure 3. Indicators of scientific consensus. Scientific consensus is formed based on a body of evidence, not a single study or investigation.



126. National Research Council, *Possible Health Effects of Exposure to Residential Electric and Magnetic Fields* (1997), <https://doi.org/10.17226/5155>; NASEM, *Cryptography and the Intelligence Community: The Future of Encryption* (2022), <https://doi.org/10.17226/26168>; NASEM, *Policing to Promote the Rule of Law and Protect the Population: An Evidence-Based Approach* (2022), <https://doi.org/10.17226/26217>.

127. World Meteorological Organization, *Scientific Assessment of Ozone Depletion: 2002*, Global Ozone Research and Monitoring Project—Report No. 47 (2003).

The path to consensus generally involves the accumulation of multiple lines of converging evidence, as studies are conducted, published, scrutinized, and iterated on, bringing more and more of the community into agreement.¹²⁸ This process works most effectively when it is carried out by a diverse group of scientists adhering to the expectations and norms outlined above. Even once broad consensus has been achieved, it is unlikely to be completely unanimous. Because scientific explanations are inherently fallible and science values skepticism, it is typical for a few in the community to remain unconvinced by even extremely compelling evidence. In fact, in a state of consensus, having a few holdouts can help maintain a degree of vigilance among the persuaded, which helps ensure the reliability of accepted knowledge. When a judge is presented with a disagreement among scientific experts, it is reasonable to seek clarification on how representative of the scientific community the two views are. Is this a case of truly unsettled science, where methodologically sound studies have been carried out and scrutinized, but open questions and disagreements remain because the hypothesis has not yet been thoroughly investigated, in which case, each party may have comparable claims on scientific reliability? Or is it a case of relatively settled science, in which one party has recruited experts who interpret the evidence differently than most of the community?¹²⁹ Note that public perception of the certainty of a scientific concept or hypothesis may differ from the actual stage of consensus building within the scientific community. This sometimes occurs as a result of strategic manipulation from stakeholders who stand to be harmed if the public were to understand the true state of scientific consensus surrounding the hypothesis, as has occurred with, for example, the health effects of tobacco, ozone depletion, and climate change.¹³⁰

Despite there being a culture of skepticism in science, consensus is often reached over time. Hypotheses are more likely to gain widespread acceptance and achieve scientific consensus if they:

128. Mary Jo Nye, *Molecular Reality: A Perspective on the Scientific Work of Jean Perrin* (1972).

129. For example, see *Allen v. Hyland's, Inc.*, which allowed testimony from experts in homeopathy, a set of rejected hypotheses about medical causation, and from a scientist noted for holding fringe views in the scientific community. See Kristin Shrader-Frechette, *Conceptual Analysis and Special-Interest Science: Toxicology and the Case of Edward Calabrese*, 177 *Synthese* 449–69 (2010), <https://doi.org/10.1007/s11229-010-9792-5>; and Jan Beyea, *Lessons to be Learned From a Contentious Challenge to Mainstream Radiobiological Science (the Linear No-Threshold Theory of Genetic Mutations)*, 154 *Env't Res.* 362–79 (2017), <https://doi.org/10.1016/j.envres.2017.01.032>.

130. For example, see Michael E. Mann, *The New Climate War: The Fight to Take Back our Planet* (2021); Naomi Oreskes, *The Scientific Consensus on Climate Change*, 306 *Science* 1686 (2004), <https://doi.org/10.1126/science.1103618>; Oreskes & Conway 2010, *supra* note 2; and G. Supran, S. Rahmsdorf & N. Oreskes, *Assessing ExxonMobil's Global Warming Projections*, *Science* 153–162 (2023), <https://doi.org/10.1126/science.abk0063>.

- have been tested and garnered support many times in many ways,
- help explain disparate and previously unexplained observations,
- can more closely explain our observations than other hypotheses,
- can be broadly applied, and
- are consistent with well-established hypotheses in neighboring fields.

Systematic Reviews and Meta-analyses

Two types of publication can be particularly useful in understanding and assessing the lines of evidence relevant to a hypothesis across multiple investigations. Systematic reviews attempt to synthesize and summarize the current state of research on a topic. In them, the authors delineate both the criteria that studies must meet for inclusion in the review and the methods that will be used to assess the studies. Meta-analyses extend this process by using statistics to analyze data from multiple studies.¹³¹ Systematic reviews and meta-analyses are used across the broad array of topics that science seeks to understand—from that which can be tightly controlled (e.g., a meta-analysis of randomized controlled trials of antimanic treatments) to complex and confounded social issues (e.g., whether school-choice programs boost student achievement).¹³² For example, in the years after Rowland and Molina published their hypothesis, scientists worked to understand how the depletion of the ozone layer might affect life on Earth. One meta-analysis examined 62 field-based studies designed to assess how an increase in ultraviolet-B radiation, which the ozone layer helps filter out, might affect plant life, concluding that the effects on plants were likely to be subtle.¹³³ Meta-analyses are an important line of scientific evidence for the courts, but, as with all scientific methodologies, require the rigorous application of reliable methodologies, as recognized in *In re Paoli R.R. Yard PCB Litigation*, in which the exclusion of a contested meta-analysis was overturned.¹³⁴

Assessing the state of scientific consensus often requires some degree of subject matter expertise and evaluation, which could be vulnerable to personal biases. Meta-analyses do not entirely sidestep these criticisms, as such analyses

131. Many guidelines for systematic reviews and meta-analyses appear in the peer-reviewed literature; however, these are generally oriented toward specific disciplines, making it difficult to cite a publication that applies broadly across the sciences.

132. Aysegül Yildiz et al., *Efficacy of Antimanic Treatments: Meta-Analysis of Randomized, Controlled Trials*, 36 *Neuropsychopharmacology* 375 (2011), <https://doi.org/10.1038/npp.2010.192>; Huriya Jabbar et al., *The Competitive Effects of School Choice on Student Achievement: A Systematic Review*, 36 *Educ. Pol'y* 247–81 (2022), <https://doi.org/10.1177/0895904819874756>.

133. Peter S. Searles, Stephen D. Flint, & Martyn M. Caldwell, *A Meta-Analysis of Plant Field Studies Simulating Stratospheric Ozone Depletion*, 127 *Oecologia* 1–10 (2001), <https://doi.org/10.1007/s0044200000592>.

134. 916 F.2d 829 (3d Cir. 1990).

are impacted by an array of assumptions and methodological approaches, about which practitioners may disagree.¹³⁵ For this and other reasons, some sociologists of science have explored approaches to assessing consensus that examine emergent properties of consensus building and that can be applied without an evaluation of the subject matter and evidence at hand, and without polling scientists or convening expert panels. For example, Shwed and Bearman examined patterns of citations within the scientific literature to assess the timing of consensus formation around the hypotheses that smoking causes cancer, that coffee does not cause cancer, and that vaccines do not cause autism.¹³⁶ This approach has also been leveraged to assess scientific consensus on the question of whether children raised by same-sex parents have similar outcomes to children raised in other family settings,¹³⁷ an analysis motivated by legal debate on the issue.¹³⁸

Not Necessarily Consensus

A few, easy-to-grasp (and so, tempting) criteria should not be taken as indicators of scientific consensus that a hypothesis is correct. These deceptive heuristics include:

- *Being called a law:* The terminology used to refer to a scientific explanation—whether hypothesis, theory, law, or model—does not indicate, in and of itself, that an idea is accurate. “Mere” hypotheses may have garnered much support and may be widely accepted, while some scientific laws have been overturned or are understood to have many exceptions. For example, Lamarck’s second law stated that traits acquired during an individual’s lifetime (e.g., muscular strength developed by lifting weights) will be passed down to one’s offspring—an idea that is at odds with most of modern genetics.
- *Statistical significance:* A statistically significant result (e.g., low *p*-value) or high confidence in a particular measurement does not itself indicate scientific consensus in the context of a single study. For example, a 2012 study found a small but statistically significant association between exposure to

135. Jop de Vrieze, *Meta-Analyses Were Supposed to End Scientific Debates. Often, They Only Cause More Controversy*, *Science* 18 (Sep. 18, 2018), <https://perma.cc/H39R-XTXN>.

136. Uri S. Shwed & Peter S. Bearman, *The Temporal Structure of Scientific Consensus Formation*, 75 *Am. Socio. Rev.* 817–40 (2010), <https://doi.org/10.1177/0003122410388488>.

137. jimi adams & Ryan Light, *Scientific Consensus, the Law, and Same Sex Parenting Outcomes*, 53 *Soc. Sci. Rsch.* 200–310 (2015), <https://dx.doi.org/10.1016/j.ssresearch.2015.06.008>.

138. In oral arguments in *Hollingsworth v. Perry*, Justice Scalia stated, “... there’s considerable disagreement among sociologists as to what the consequences of raising a child in a single-sex family, whether that is harmful to the child or not.” (Transcript of oral argument at 19, *Hollingsworth v. Perry*, 570 U.S. 693 (2013), https://www.supremecourt.gov/oral_arguments/argument_transcripts/2012/12-144_5if6.pdf).

Zoloft in utero and major congenital malformations.¹³⁹ But other studies have reached different conclusions on this question, and a later meta-analysis that included the 2012 study found no significantly increased risk for major congenital anomalies associated with Zoloft use during pregnancy.¹⁴⁰ Despite an earlier statistically significant result, scientific evidence appears to be converging around there being no or minimal elevated risk for these birth defects as a result of a pregnant person's Zoloft use. While one's confidence in the conclusions of a lone study should always be tempered, if a particular method, measure, or analytic technique has been established over the course of multiple studies and is widely accepted by the scientific community, then one might well feel highly confident about the reliability of a significant result obtained through a particular application of that approach, bearing in mind that statistically significant results necessarily occur by chance alone some of the time.

- *Peer-reviewed publication:* This is an important milestone along the road to achieving consensus, but of course, many hypotheses published in peer-reviewed journals are ultimately rejected by the scientific community and do not make it very far in that journey. For example, before Watson and Crick published their ideas about the structure of DNA, a peer-reviewed publication proposed the now-rejected hypothesis that DNA is a three-stranded molecule.¹⁴¹
- *Journal prestige, scientific eminence, or high citation metrics alone:* Hypotheses published in journals with high standards for evidence and stringent peer-review practices that then go on to receive many citations because other scientists are building on the research is an indication that the scientific community is moving toward broad acceptance of a hypothesis; however, any one of these factors in isolation is a poor indicator of consensus. The fact that an article has high citation metrics, was published in a prestigious journal (often loosely indicated by the journal's impact factor), or has a famous author is not by itself a reason to conclude that the study's findings are widely accepted scientific facts. The vagaries, inconsistencies, and biases of publishing are discussed above, and plenty of renowned scientists have been wrong in prestigious journals. The proposal for the three-stranded DNA helix was authored by Linus Pauling, winner of the Nobel Prize in Chemistry, in the *Proceedings of the*

139. Espen Jimenez-Solen et al., *Exposure to Selective Serotonin Reuptake Inhibitors and the Risk of Congenital Malformations: A Nationwide Cohort Study*, 2 *BMJ Open* e001148 (2012), <https://doi.org/10.1136/bmjopen-2012-001148>.

140. Shan-Yan Gao et al., *Selective Serotonin Reuptake Inhibitor Use During Early Pregnancy and Congenital Malformations: A Systematic Review and Meta-Analysis of Cohort Studies of More than 9 Million Births*, 16 *BMC Med.* 205 (2018), <https://doi.org/10.1186/s12916-018-1193-5>.

141. Linus Pauling & Robert B. Corey, *A Proposed Structure for the Nucleic Acids*, 39 *Proc. of the Nat'l Acad. of Sci. USA* 84–97 (1953), <https://doi.org/10.1073/pnas.39.2.84>.

National Academy of Sciences, for example. Furthermore, women, people of color, other historically oppressed groups, and non-Western people have been, and in many cases continue to be, unfairly dismissed and so may lack the recognition their work deserves, which includes prestigious appointments and publications.¹⁴² Ultimately, evidence, not reputation, determines scientific acceptance, although these other factors are likely to influence how long it takes the community to reach consensus and who receives credit for developing the knowledge.

Because scientific consensus emerges from community-level processes driven by scrutiny, additional testing, and application of a hypothesis, the simple indicators listed here are not, in isolation, signals of consensus.

Myths About Science

Above we described and debunked some common myths about science, which include the following.

Myth: There is a single scientific method that all scientists follow.

Fact: The process of science is nonlinear and dynamic.

Myth: The institution of peer review ensures that all published papers are sound and dependable.

Fact: The conclusions that many peer-reviewed articles reach will ultimately turn out to be incorrect.

Myth: Without an experiment, a scientific investigation cannot be rigorous or reliable.

Fact: Nonexperimental evidence is critical in evaluating many scientific ideas.

Myth: Qualitative data are “soft” and relatively unimportant.

Fact: Qualitative data are essential to providing context and meaning to quantitative results, especially within social science, and are key to generating hypotheses in many disciplines.

A few other common misconceptions about science warrant unpacking.

Myth: Hard sciences are more rigorous and scientific than so-called soft sciences.

142. See Esther A. Odekunle, *Dismantling Systemic Racism in Science*, 369 *Science* 780–81 (2020), <https://doi.org/10.1126/science.abd7531> and references therein.

How Science Works

Fact: Both the classic hard sciences (physics, chemistry, biology, geology, and astronomy) and the social sciences, which are sometimes portrayed as soft, can be approached with well-designed, rigorous investigations and develop reliable explanations for phenomena.¹⁴³

As later reference guides in this manual make clear, many questions dealing with economics, decision making, and human psychology are legitimate topics of scientific investigation and employ rigorous, reliable methodologies. A variety of experimental, nonexperimental, qualitative, and quantitative methodologies are employed by all of these fields, as well as in the traditional natural sciences, and judges can apply similar considerations regarding the process of science, as outlined in this reference guide, to all of them. Of course, the more complex a study system is, the more difficult it is to control variables and untangle causal mechanisms. Furthermore, tightly controlling a study system can make it more difficult to apply those results to real-world problems. Many systems of concern in the social sciences, earth sciences, and biology are enormously complex, posing methodological challenges; however, because scientific questions in these fields are also of pressing importance to human well-being, scientists go to great lengths to meet and overcome such challenges.

Myth: Science undergoes periodic scientific revolutions, or paradigm shifts, in which an old understanding of a natural system is replaced by a new one, and the two provide such different worldviews that it is not possible to make comparisons of evidence, or perhaps even communicate, across the two perspectives.

Fact: Science moves forward incrementally, as well as in large leaps. These changes occur because evidence has accumulated, and the new view makes sense of more evidence more coherently than did the old one.

Philosopher Thomas Kuhn famously conceptualized large-scale changes in accepted scientific knowledge as paradigm shifts.¹⁴⁴ According to Kuhn, the history of science can be divided up into times of normal science, when scientists add to and elaborate on an accepted scientific theory, such as Newton's classical mechanics, and briefer periods of revolutionary science, when there is a switch to a new theory or paradigm—in this case, Einstein's theory of relativity, which explains everything that classical mechanics did and more. Kuhn argued that new and old paradigms were incommensurable, though later retreated from this view somewhat. Philosophers and historians of science generally agree that many details of Kuhn's concept of paradigm shifts do not align well with how changes

143. Larry V. Hedges, *How Hard Is Hard Science, How Soft Is Soft Science? The Empirical Cumulativeness of Research*, 42 *Am. Psych.* 443–55 (1987), <https://doi.org/10.1037/0003-066X.42.5.443>.

144. Thomas S. Kuhn, *The Structure of Scientific Revolutions* (1962).

in scientific explanations actually happened.¹⁴⁵ While Kuhn's ideas are not a broadly accepted account of scientific change among modern philosophers, sociologists, and historians of science, they were influential in many ways, for example by focusing attention on theory change within science and the potential role of the resolution of evidentiary anomalies in this process.

Myth: Science moves forward only by falsifying, rejecting, or disproving hypotheses, leaving the “last hypothesis standing” as the most likely to be correct.

Fact: Science can neither prove nor disprove hypotheses, and scientists evaluate hypotheses based on both supporting and refuting evidence.

This misconception is based on the idea of falsification, philosopher Karl Popper's influential account of scientific justification, which suggests that all science can do is reject, or falsify, hypotheses and that science cannot find evidence that *supports* one idea over others.¹⁴⁶ Falsification is a popular philosophical doctrine, especially with scientists, and apparently with judges as well, as the *Daubert* court explained, “Scientific methodology today is based on generating hypotheses and testing them to see if they can be *falsified* [emphasis added]; indeed, this methodology is what distinguishes science from other fields of human inquiry.”¹⁴⁷ However, falsification is not a complete or accurate picture of how scientific knowledge is built. First of all, reviewing the logical arguments presented in any scientific journal will reveal that evidence can and does play a role in *supporting* particular hypotheses over others, not just in ruling some ideas out, as implied by the doctrine of falsification. Furthermore, philosophers of science now recognize that science cannot once-and-for-all, absolutely *prove* any idea to be false or true.¹⁴⁸ Even if all the available evidence lines up for or against a particular hypothesis, science always allows for the outside possibility that we are missing something—for example, that one of our assumptions about the system is incorrect and so our tests are not measuring what we think they are, that our interpretation of the evidence is shaded by societal norms and that future generations will see the results differently, or that there is a subtle bug in the code of a common statistical package that is leading everyone in the same wrong direction. Those are certainly very, very unlikely possibilities, but they are not strictly impossible; hence, the essential fallibility of scientific knowledge. Of course, this does not

145. Kitcher 1995, *supra* note 5.

146. Karl R. Popper, *Conjectures and Refutations: The Growth of Scientific Knowledge* (1963).

147. See *Daubert*, 509 U.S. 579 at 593 (citing Michael D. Green, *Expert Witnesses and Sufficiency of Evidence in Toxic Substances Litigation: The Legacy of Agent Orange and Bendectin Litigation*, 86 Nw. U. L. Rev. 643 (1992)).

148. Godfrey-Smith 2003, *supra* note 3; Kitcher 1983, *supra* note 27; and Kitcher 1995, *supra* note 5.

mean that all scientific hypotheses are equally likely; in fact, it's the opposite. The whole business of science is sorting through explanations and collecting evidence to try to find those that are most expansive and powerful and are supported by the widest variety of evidence. Despite the fact that scientists can never reach absolute proof or disproof, there are plenty of hypotheses that scientists are remarkably certain are either true or untrue, and these ideas are extraordinarily unlikely to ever be overturned. Nonscientists justifiably place their trust in this same set of reliable, accepted hypotheses every time we rely on the myriad of technologies, policies, and other applications that derive from them.

Myth: If a scientific idea is called a *hypothesis* or a *theory*, this indicates that the explanation is uncertain, whereas *laws* are nearly irrefutable scientific ideas.¹⁴⁹

Fact: Hypotheses, theories, and laws are scientific explanations that differ in breadth, not in level of support, and these terms are not consistently used across scientific disciplines or throughout history.

In everyday language, the words *hypothesis* and *theory* usually refer to educated guesses or ideas that we are quite uncertain about, while *laws* are viewed as rigid mechanisms that have few exceptions. However, the meaning of these words differs within modern science, where *hypotheses* are generally understood to be potential explanations for natural phenomena regardless of the extent to which the explanations have been investigated or the amount of evidence supporting or refuting them; the term *law* is often used to refer to a statement (often a mathematical statement) about how observable phenomena are related; and *theories* are understood to be deep explanations that apply to a broad range of phenomena and that may integrate many hypotheses and laws. Any individual hypothesis, theory, or law may have accumulated strong evidence supporting or refuting it—or little in either direction. Adding further confusion, use of these terms is not standardized, and they have gained popularity in different scientific disciplines and at different points in history. Thus, we have Newton's *laws*, which are still widely applicable but were subsumed at an explanatory level by Einstein's now accepted *theory* of relativity; Lamarck's *laws*, now rejected, their explanatory position occupied by the *theory* of evolution; and the once-heretical prion *hypothesis*, now settled science and the basis for the 1997 Nobel Prize for Medicine or Physiology. In practice, judges should set little store by the terminology used to refer to a unit of purported scientific knowledge and focus instead on the methodologies and evidence that underlie it.

Myth: Scientists are completely objective in their evaluation of scientific ideas and evidence.

149. McComas 1998, *supra* note 4.

Fact: Scientists' work is influenced by their biases, motivations, backgrounds, and experiences.

Scientists may be motivated to different degrees by curiosity, by the desire to solve a problem or gain recognition or funding, or by personal financial interests. These factors and many other conscious and unconscious biases shape the questions scientists ask, how they investigate these questions, how they interpret evidence, and how they evaluate and credit the work of others. However, because scientists are expected to, incentivized to, and generally do strive for objectivity and because of the ongoing community scrutiny of scientific findings from many different viewpoints, such biases are attenuated and ultimately corrected—but this does take time. In the context of a trial concerning as-yet unsettled science, judges can reasonably weigh potential sources of bias, particularly when research has shown this to influence results, as in the case of funding bias.

Myth: Research that comes out of universities is pure and unbiased by financial and other considerations.

Fact: All research is vulnerable to bias.

The same arguments about bias that apply to individual scientists also apply to research institutions of all sorts. Universities, government agencies, organizations, and industries all have complex internal and external incentive systems that shape how they operate and carry out research. Universities may earn revenues from patents, and many have offices devoted to handling intellectual property. Such conflicts of interest can be relevant in assessing potential sources of bias at trial.

Science and the Law

Science and the law are sometimes seen to be clashing cultures, but in fact they have many overlapping properties and interests.¹⁵⁰ Both are oriented toward discovering truths and facts. Both are complex endeavors, involving many interacting individuals and institutions that, while guided by overarching ideals, are nonetheless embedded in and shaped by the cultures and values of society and of their institutions. Both endeavors are influenced by the backgrounds, experiences, biases, and values of their human practitioners but have mechanisms and guidelines designed to help them achieve their aims and identify truths. Here, we outline further similarities and differences between science and the law that may help judges bridge these two cultures and make useful connections between them.

150. Jasanoff 2005, *supra* note 5.

Assessment of Evidence

Both science and law rely on evidence. Judges use scientific evidence to strengthen their ability to understand the factual basis of a dispute to reach a fair and just resolution of a conflict. Judges must make preliminary assessments of scientific information presented as evidence at trial to ensure that information meets appropriate standards of scientific integrity for consideration by a jury. Scientists, on the other hand, use evidence to make judgments about the accuracy of hypotheses.

Both institutions put individuals in the position of weighing the often-incomplete evidence to come to a conclusion about what the likely truth is. In the courtroom, judges and juries must make decisions based on the current state of the scientific evidence presented during a trial. This distinct terminus to deliberation is a requirement of the law but is not generally a part of science. In most cases, scientists can and will reserve judgment on a hypothesis if warranted, waiting for more evidence before accepting or rejecting it. This difference between science and the law was acknowledged in the decision on the Zoloft litigation referenced above: “The Court recognizes that the final scientific verdict as to whether Zoloft can cause birth defects may not be delivered for many years. Nevertheless, Plaintiffs chose when to file their cases, and the Court concludes that for the Plaintiffs who have continued to pursue their claims, the litigation gates must be closed.”¹⁵¹

Within the law, the standards for assessing evidence vary by jurisdiction and type of action. Criminal cases, for example, require proof beyond a reasonable doubt to draw the conclusion of wrongdoing, while civil cases require only a preponderance of evidence. As described in *The Admissibility of Expert Testimony*, in this manual, scientific evidence is admissible if the judge finds it “more likely than not that the expert’s methods are reliable and that they are reliably applied to the facts at hand.” Scientists too may vary their standards for evidence depending on the situation and the implications of the conclusion, but these recalibrations are not standardized across discipline or situation, as is the case in litigation. Scientists may, for example, lower their threshold for action when the investigators perceive a threat of harm, as occurred when Molina and Rowland began lobbying for policy change long before scientific consensus around the connection between CFCs and ozone depletion was achieved.

In the United States, our common-law tradition often relies on a lay jury to assess evidence to resolve disputes related to science. However, in science, only in rare cases is a specific body tasked with assessing the evidence and determining scientific consensus (see section titled “Achieving Scientific Consensus” above). Instead, consensus is usually an emergent property of interactions within the scientific community. Within science, disputes about the accuracy of

151. *In re Zoloft* (Sertraline Hydrochloride) Prods. Liab. Litig., 176 F. Supp. 3d 483 (E.D. Pa. 2016).

hypotheses are managed by experts in that field and ultimately depend on the evidence, not the lay public's perception of the issues. For example, scientific experts have long since reached consensus on the fact that measles, mumps, and rubella (MMR) vaccines do not cause autism; however, this concern still looms large for some of the lay public, continues to make its way into the courts, and may find sympathies within the jury box.

The emphasis in the legal system on resolving legal conflicts within the common-law tradition may require presentation of scientific testimony in ways that are uncommon in the ordinary course of science. For example, the legal system allows parties or the court to select the scientific experts who will testify, advise, or mediate, and parties involved with the case may identify experts they perceive to be sympathetic to their perspective, rather than seeking out the most knowledgeable scientists or scientists whose interpretation of the evidence represents that of most of the relevant scientific community. In addition, in litigation, attorneys and judges determine how and what scientific evidence is presented, potentially limiting the jury's access to certain information. In contrast, interactions within the scientific community are not tightly regulated in terms of who is permitted to present and assess what evidence and in what ways, although journal editors and peer reviewers do play a role in determining what scientific evidence is available for review through publications. In science, debates relating to scientific evidence generally play out over many years, most notably by interactions among any subject matter experts who opt into the discussion, without imposed limits regarding what information can be used to buttress or refute an argument.

Precedent and Self-Correction

In the United States, precedent is built into our legal system. According to *stare decisis*, cases with the same facts are expected to be decided in the same way. Though precedent can be overturned, the hierarchical structure of the judiciary makes this occurrence rarer than it would be otherwise, and precedence remains a powerful doctrine in the law. Precedence has a limited role in science, on the other hand. Hypotheses are judged according to the specific evidence relevant to them, and scientists are expected to give up a previously accepted hypothesis if new evidence or interpretations warrant it. This is considered a normal part of good science, and science has built in mechanisms that allow for self-correction. Science is, of course, done by people, who may cling to ideas or practices, in institutions with cultures that may be slow to change, so science *does* experience a form of precedent. But because this is not a formal mechanism of science, and indeed conflicts with the idealized modes of reasoning that scientists expect of themselves, precedence is less important in science than in the law. This

difference can contribute to a lag time between a shift in scientific consensus and when this consensus is consistently applied to proceedings in courts. For example, bitemark evidence has little methodologically sound science to support its use (see the *Reference Guide on Forensic Feature Comparison Evidence*, in this manual), a fact that started to become clear in the early 2000s.¹⁵² Yet, bitemark evidence continues to be admitted in some courts.

Conclusion

We've presented science as a human endeavor emerging out of a community that has tasked itself with working together to answer a wide variety of questions about the natural and human world through an iterative, self-correcting process. While science is often slow and its hypotheses fundamentally tentative, it is nonetheless a reliable means of generating explanations. On the broadest scale, science helps us understand the way the world is, connects disparate phenomena, and makes accurate and consistent predictions. At a practical level, science helps us solve problems, develop new technologies, make decisions, form policies, and, as this manual illustrates, administer justice. Subsequent reference guides in this manual provide a road map to the scientific findings that are most likely to factor into judicial decisions, guiding judges and juries in their search for truth and in their service to the law.

152. Michael J. Saks et al., *Forensic Bitemark Identification: Weak Foundations, Exaggerated Claims*, 3 J.L. & Biosciences 538–75 (2016), <https://doi.org/10.1093/jlb/lsw045>.

Glossary of Terms

- auxiliary hypothesis.** An assumption made in the context of testing a main hypothesis.
- blinding.** The concealment of comparison group membership so that this knowledge does not bias the study outcome.
- control group.** A condition or group that does not receive the experimental manipulation, which serves to increase confidence that the change observed in the experimental group is caused by the manipulation or factor in question.
- controlled experiment.** An experiment that seeks to keep variables other than the experimental manipulation the same to help isolate the cause of any change.
- effect size.** A statistic that communicates the strength of an association between variables.
- error.** In reference to statistics, the potential difference between a computed or measured value and the true value.
- experiment.** A testing methodology that involves intentionally manipulating some factor in a system to learn how that affects the outcome.
- fallibilism.** The principle that holds that scientific knowledge is always open to rejection or revision, no matter how much prior evidence supports it.
- falsification.** A now outdated account of scientific justification, which proposes that science can only prove hypotheses wrong and that scientific acceptance arises through a process of elimination.
- hypothesis.** Within science, a potential explanation for a natural phenomenon, regardless of the extent to which the explanation has been investigated or the amount of evidence supporting or refuting it, although this term is inconsistently applied and its use is more or less common in different disciplines and at different points in history.
- law.** Within science, an often-mathematical statement of the relationship among observable phenomena, although this term is inconsistently applied and its use is more or less common in different disciplines and at different points in history.
- meta-analysis.** A study that uses statistics to analyze data from multiple other studies, aggregating and synthesizing their results.
- mixed methods.** A research approach that uses a combination of qualitative and quantitative techniques within a single investigation to detect patterns and to form and test hypotheses.
- model.** Usually, a mathematical representation of hypothesized mechanisms and interactions within a system that can be used to investigate the impact

of parameter changes on predicted system outcomes, although this term is inconsistently applied and its use is more or less common in different disciplines and at different points in history.

natural. Concerning any element of the physical universe, whether made by humans or not.

peer review. In science, a standardized process in which journal articles (or other scientific communications, such as funding applications) are vetted by other subject matter experts before acceptance or publication.

placebo. In medical studies, a treatment, medicine, or therapy given to study participants that is not known to have a therapeutic effect on the condition of interest.

power. A measure of a statistical test's ability to detect an effect that is present.

prediction. A potential outcome of a scientific test that is arrived at by logically reasoning about what we would expect to observe if a hypothesis were true or false.

preprint. A fully drafted scientific journal article that includes all the features of a published article but that has not yet been accepted to a journal for publication.

***p*-value.** A statistic that gives the calculated probability that the null hypothesis could be true even given the observed differences between conditions.

randomized controlled trial. An experimental design that randomly assigns participants to the experimental or control group to better control variables across the groups.

replication. Herein, a study carried out with the intent of mimicking another to the closest degree possible.

replication crisis. A series of observations, beginning around 2010, suggesting that the results of many published scientific studies are not replicable and their conclusions potentially incorrect.

science. A body of knowledge regarding the natural world and the process for building that knowledge based on evidence acquired through observation, experiment, and simulation.

systematic review. A publication type that attempts to synthesize and summarize the current state of research on a topic.

theory. Within science, a concise, coherent, and predictive explanation for a broad range of natural phenomena that integrates and makes sense of many hypotheses, although this term is inconsistently applied and its use is more or less common in different disciplines and at different points in history.

uncertainty. In reference to statistics, a parameter that indicates the dispersion of values within which the true value is likely to fall.

Reference Guide on Forensic Feature Comparison Evidence

VALENA E. BEETY, JANE CAMPBELL MORIARTY, AND
ANDREA L. ROTH

Valena E. Beety, Robert H. McKinney Professor of Law, Indiana University Maurer School of Law.

Jane Campbell Moriarty, Carol Los Mansmann Chair in Faculty Scholarship and Professor of Law, Thomas R. Kline School of Law at Duquesne University.

Andrea L. Roth, Professor of Law and Barry Tarlow Chancellor's Chair in Criminal Justice, UC Berkeley School of Law.

CONTENTS

Overview, 115

Legal Framework Governing Forensic Feature Comparison Evidence, 116

 Prior Government and NGO Review of Forensic Feature
 Comparison Evidence, 122

 Recurring Concerns with Forensic Feature Comparison
 Evidence, 129

 Terminology and Testimony Issues, 129

 Reliability Concerns, 134

Discussion of Selected Forensic Feature Comparison Evidence, 140

 Differences Between DNA and Other Methods, 140

 Fingerprint Evidence, 141

 Introduction, 141

 The Method and Its Claims, 142

 Scientific Assessments, Critiques, and Debates, 144

 Case Law Development, 151

 Handwriting Evidence, 153

 Introduction, 153

 The Method and Its Claims, 154

 Scientific Assessments, Critiques, and Debates, 157

 Case Law Development, 160

 Firearms and Toolmark Analysis, 165

 Introduction, 165

 The Firearms Comparison Method and Its Claims, 168

 The Toolmark Comparison Method and Its Claims, 171

 Scientific Assessments, Critiques, and Debates, 172

 Case Law Development, 178

Reference Manual on Scientific Evidence

- Bitemark Evidence, 181
 - Introduction, 181
 - The Method and Its Claims, 182
 - Scientific Assessments, Critiques, and Debates, 185
 - Case Law Development, 187
- Facial Recognition Software and Other Methods, 189
- Special Issues with Machine-Generated Feature Comparison Evidence, 192
 - Types of Machine-Generated Feature Comparison Evidence, 192
- Recurring Legal Issues, 195
 - Rule 702/Reliability Issues, 195
 - Hearsay and Confrontation, 197
 - Discovery Issues, 199
- Glossary of Terms, 202

Overview

This reference guide discusses techniques used for forensic feature comparison, meaning comparing features in trace evidence (such as a latent fingerprint, handwriting on a document, or toolmarks on a projectile) to a reference sample (such as a suspect's reference fingerprints or handwriting exemplar, or a projectile fired from a known weapon) to determine whether the samples might originate from the same source. These techniques play a significant role in modern federal litigation, particularly criminal cases. In turn, federal judges play a central gatekeeping and managerial role in regulating the admissibility and use at trial of these techniques. The role is inherently challenging, as judges are asked to weigh in on complex issues of scientific validity that are the realm of experts. The role is also particularly challenging now, as some techniques' ability to accurately determine whether two patterns have a common source have been challenged as unreliable in light of high-profile DNA exonerations and recent governmental reports critical of feature comparison disciplines. Meanwhile, the increasing automation of feature comparison techniques with the help of software raises unique issues.

With these realities in mind, the goal of this reference guide is to provide an accessible overview for judges as they resolve legal questions about forensic feature comparison evidence. While other guides in this manual cover in depth the legal rules related to expert testimony and the specific issues related to forensic DNA typing, this reference guide offers judges a solid background in scientific and legal issues related to a number of forensic feature comparison techniques. This guide's scope is limited to feature comparison techniques, as such techniques raise recurring issues worthy of separate treatment and do not include other technical or scientific fields such as medicolegal death investigations, seized drug analysis, or digital forensics.

The first section of this reference guide explains the basic legal framework for determining the admissibility of forensic feature comparison evidence, situating the case law and rules discussed in *The Admissibility of Expert Testimony*, in this manual, in the specific context of this type of evidence.

The second section catalogs the most frequently cited institutional efforts to evaluate forensic feature comparison evidence, such as the 2009 National Research Council report (2009 NRC Report), the 2016 report of the President's Council of Advisors on Science and Technology (PCAST), the National Commission on Forensic Science (NCFS), and the National Institute of Standards and Technology (NIST)'s Organization of Scientific Area Committees (OSAC). We catalog the major findings of these groups not only because they are frequently cited by litigants but because many lower-court rulings on the admissibility of forensic feature comparison evidence were issued before these reports and may not be consistent with them.

The third section of this reference guide explains the continuing concerns that feature comparison evidence poses, with respect to experts' arguable

overclaims of certainty or individualization, frequently confused terms such as reliability versus validity, class versus individual characteristics, and other matters. This section also provides an in-depth discussion of foundational reliability and reliability as applied.

The fourth section offers a detailed explanation of some of the forensic feature comparison techniques currently in use and most likely to arise in federal cases in the next decade. These techniques include fingerprint analysis, handwriting comparison, firearms and toolmark identification, and bitemark evidence. While some techniques like bitemark analysis and handwriting comparison appear to be gradually receding in use, we include them both because they are still offered as proof and because they are sure to arise in habeas litigation or other postconviction proceedings for years to come. This guide omits several other techniques, such as shoeprint comparisons, hair and fiber analysis, and glass comparisons, that present some of the same issues as those techniques that are covered.¹ Each overview explains the basic method, the extent of empirical studies on the method, and the method's legal status in the courts. We end with a brief discussion of facial recognition and other emerging techniques.

Finally, this reference guide offers a section on automated, machine-generated feature comparison conclusions. While the *Reference Guide on Human DNA Identification Evidence*, in this manual, discusses DNA software, other feature comparison techniques are also being increasingly automated. We offer an overview of the types of machine-generated forensic conclusions a federal judge is likely to see, as well as the legal issues likely to arise in admitting these conclusions. The goal of this final section is not to definitively resolve these issues, but to equip judges to understand the stakes and assumptions underlying the conflicts they themselves will need to resolve.

Legal Framework Governing Forensic Feature Comparison Evidence

This section builds on *The Admissibility of Expert Testimony*, in this manual, by briefly highlighting significant aspects of Federal Rule of Evidence 702's application to forensic feature comparison evidence in particular. As *The Admissibility of Expert Testimony* discusses in depth, Rule 702 codifies the so-called “*Daubert* trilogy” and requires that an expert be qualified, that the opinion be helpful to the jury, that the testimony be based on sufficient facts or data, and that the testimony is the product of a method that is both foundationally reliable and has been reliably

1. Judges can find a more comprehensive list of feature comparison disciplines by perusing OSAC's website: <https://perma.cc/8W8P-DHFT>. Clicking on the link of each subcommittee leads to a list of proposed standards on a host of subdisciplines.

applied to the facts of the case.² As *Daubert* noted, courts determining the reliability of scientific expert testimony look to various factors, such as whether a method has been tested; the method's error rate; whether the method has been subject to peer review; whether the method is governed by standards and protocols; and whether the method is generally accepted in the relevant scientific community.³

As noted in *The Admissibility of Expert Testimony*, the 2023 amendments to Rule 702 and the advisory committee's note make explicit that the proponent must meet the requirements above by a preponderance of the evidence.⁴ The advisory committee's note urges special caution with "subjective" forensic methods:

The amendment is especially pertinent to the testimony of forensic experts in both criminal and civil cases. Forensic experts should avoid assertions of absolute or one hundred percent certainty—or to a reasonable degree of scientific certainty—if the methodology is subjective and thus potentially subject to error. In deciding whether to admit forensic expert testimony, the judge should (where possible) receive an estimate of the known or potential rate of error of the methodology employed, based (where appropriate) on studies that reflect how often the method produces accurate results. Expert opinion testimony regarding the weight of feature comparison evidence (i.e., evidence that a set of features corresponds between two examined items) must be limited to those inferences that can reasonably be drawn from a reliable application of the principles and methods.⁵

What follows is a brief overview of each admissibility requirement in the context of feature comparison evidence.

Helpfulness to the jury. First, the skill of non-DNA feature comparison experts may well be "helpful to the jury" in explaining the evidence. For example, in a firearm comparison case, one federal district court recently stated that the expert's "technical knowledge, skill, and training to make microscopic observations of and comparisons between cartridge cases, would be helpful to the trier of fact."⁶ Courts have come to similar conclusions with respect to other feature comparison disciplines.⁷ But these rulings presuppose that the conclusion itself is reliable and thus worthy of being heard by the jury. In short, "helpfulness to the jury" is typically not the dispositive factor affecting admissibility of feature

2. Fed. R. Evid. 702.

3. See *Daubert v. Merrell Dow Pharms., Inc.*, 509 U.S. 579 (1993) (listing nonexhaustive factors in determining reliability of a scientific expert method).

4. See Fed. R. Evid. 702 & advisory committee's note to 2023 amendment.

5. See Fed. R. Evid. 702 advisory committee's note to 2023 amendment.

6. *United States v. Davis*, No. 4:18-CR-00011, 2019 WL 4306971, at *6 (W.D. Va. Sept. 11, 2019).

7. See, e.g., *United States v. Dale*, 618 F. App'x 494 (11th Cir. 2015) (fingerprint and handwriting comparison); *United States v. Shipp*, 422 F. Supp. 3d 762, 783 (E.D.N.Y. 2019) (firearm comparison); *United States v. Mallory*, 902 F.3d 584, 593 (6th Cir. 2018) (handwriting comparison).

comparison discipline evidence once it is deemed foundationally reliable and reliable as applied.

Expert qualifications. When allowing a feature comparison discipline expert to testify “before the jury cloaked with the mantle of an expert,” a trial court must “assure that a proffered witness truly qualifies as an expert.”⁸ Trial courts should be careful not to conflate expert qualification with either foundational reliability or reliability as applied. A method might be foundationally reliable, but wielded by an examiner who is not sufficiently proficient to be qualified as an expert to begin with. In turn, as explained more fully below, even an expert with established credentials and proficiency must demonstrate through documentation that the method was reliably applied in the case at hand.

In theory, Rule 702 allows much flexibility to the proponent in how to show an expert is qualified. The rule is written in the disjunctive—qualified by “knowledge, skill, experience, training, or education”—and provides multiple avenues for qualifying witnesses, including experience alone.⁹ Moreover, experts need not be “blue-ribbon practitioners” with “optimal qualification[s],”¹⁰ or even “highly qualified in order to testify about a given issue.”¹¹

Still, courts must provide sufficient reasons for finding the expert qualified so that an appellate court can determine “whether the district court properly applied the relevant law.”¹² The question is not merely whether the witness is qualified in a vacuum to speak on expert matters; it is whether the witness is qualified to offer the opinion she is offering. Some courts have excluded feature-comparison experts on this ground.¹³ An expert might also be qualified to render opinions on some topics but not others.¹⁴ Specifically, some (but not all) courts have deemed

8. *United States v. Cloud*, No. 1:19-CR-02032-SMJ-1, 2022 WL 801694, at *1 (E.D. Wash. Mar. 8, 2022) (fingerprint expert) (quoting *Jinro Am. Inc. v. Secure Invs., Inc.*, 266 F.3d 993, 1004 (9th Cir. 2001)).

9. The advisory committee’s note to the 2000 amendment of Rule 702 states that “[n]othing in this amendment is intended to suggest that experience alone . . . may not provide a sufficient foundation for expert testimony.” Fed. R. Evid. 702 advisory committee’s note to 2000 amendment.

10. *United States v. Vargas*, 471 F.3d 255, 262 (1st Cir. 2006).

11. *Huss v. Gayden*, 571 F.3d 442, 452 (5th Cir. 2009).

12. *United States v. Avitia-Guillen*, 680 F.3d 1253, 1259 (10th Cir. 2012) (no abuse of discretion ruling that fingerprint expert was qualified; district court provided sufficient information to establish basis for the opinion).

13. *See, e.g., Balimunkwe v. Bank of Am.*, No. 1:14-CV-327, 2015 WL 5167632 (S.D. Ohio Sept. 3, 2015) (expert not permitted to testify where he did not possess sufficient qualifications as handwriting comparison expert); *Almeciga v. Ctr. for Investigative Reporting, Inc.*, 185 F. Supp. 3d 401, 424 (S.D.N.Y. 2016) (same).

14. *See, e.g., Marten Transp., Ltd. v. Plattform Advertising, Inc.*, 184 F. Supp. 3d 1006 (D. Kan. 2016) (holding that expert witness, though qualified to render several opinions on nature of trucking industry and hiring practices, was not qualified to render opinion on search engine optimization).

the exclusion of experts such as law professors, who themselves are not scientists but who have written extensively about a discipline, not an abuse of discretion.¹⁵

Before the 2023 amendments to Rule 702, some courts had come close to deeming a forensic feature comparison expert unqualified, but concluded that qualification was a low enough bar that any questions should go to weight, not admissibility.¹⁶ The 2023 amendments explicitly note that, while questions of precise weight as to the opinions of a qualified expert are for the jury, the issue of qualification (as well as helpfulness and reliability) goes to admissibility and not merely weight.¹⁷ Ultimately, appellate courts reviewing qualifications of forensic feature or pattern comparison experts have affirmed that it is the judge's obligation, and not a jury question, to determine whether a witness possesses sufficient qualifications.¹⁸

Some recent commentary argues that, given the 2023 amendments' concern over "subjective" methods based on examiner judgment and experience, trial courts should take proficiency testing results (or lack thereof) of feature comparison experts more seriously at the expert qualification stage.¹⁹ Several consensus reports and other scholars have similarly urged courts to be more careful gatekeepers about the qualifications of feature comparison experts, noting that the proficiency tests that exist to date are not encouraging and the results may be misleading, as the tests do not simulate realistic crime-scene samples.²⁰

15. See, e.g., *State v. Clifford*, 121 P.3d 489 (Mont. 2005) (holding that trial court did not abuse its discretion in excluding law professor as expert in handwriting); *United States v. Paul*, 175 F.3d 906, 912 (11th Cir. 1999) (same).

16. See, e.g., *Banuchi v. City of Homestead*, 606 F. Supp. 3d 1262, 1272–73 (S.D. Fla. 2022) (finding an expert "barely" qualified to testify about why fingerprints would not be at a crime scene, reasoning that "[t]he qualification standard for expert testimony is not stringent, and so long as the expert is *minimally* qualified, objections to the level of the expert's expertise [go] to credibility and weight, not admissibility") (quoting *Vision I Homeowners Ass'n, Inc. v. Aspen Specialty Ins. Co.*, 674 F. Supp. 2d 1321, 1325 (S.D. Fla. 2009) (emphasis added) (quoting *Kilpatrick v. Breg, Inc.*, Case No. 08-10052-CIV, 2009 WL 2058384, at *3 (S.D. Fla. June 25, 2009))); *Thomas v. United States*, No. 2:03-CV-02416-JPM, 2015 WL 5076969, at *184 (W.D. Tenn. Aug. 27, 2015), *aff'd*, 849 F.3d 669 (6th Cir. 2017) (explaining the witness's lack of qualifications in the area of handwriting comparison, but admitting it with the caveat that it was entitled to "little weight" in a bench trial).

17. See Fed. R. Evid. 702 & advisory committee's note to 2023 amendment.

18. See, e.g., *United States v. Ruvalcaba-Garcia*, 923 F.3d 1183, 1189 (9th Cir. 2019) (trial judge erred in ruling that fingerprint technician's qualification "as an expert" was a "determination for the jury").

19. See, e.g., Brandon L. Garrett & Gregory Mitchell, *The Proficiency of Experts*, 166 U. Pa. L. Rev. 901, 902 (2018) (explaining their evaluation of twenty years of fingerprint proficiency tests reflected a "surprisingly high rates of false positive identifications" and that "credentials and experience are often poor proxies for proficiency").

20. See, e.g., President's Council of Advisors on Sci. & Tech. (PCAST), Exec. Office of the President, *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods* 51 (Sept. 2016), <https://perma.cc/GJ9A-2DFS> [hereinafter 2016 PCAST Report]; Garrett & Mitchell, *supra* note 19, at 902 (explaining their evaluation of twenty years of fingerprint proficiency tests reflected a "surprisingly high rates of false positive identifications");

Foundational reliability. This section provides a brief reminder of the foundational reliability requirements of Rule 702. A more in-depth discussion of specific recurring reliability issues in feature comparison evidence is set forth in the next section.

In determining foundational reliability of feature comparison expert testimony, judges have multiple options under Rule 702's framework, consistent with their role as gatekeeper: (1) they may decide that the evidence is admissible; (2) they may exclude the evidence entirely, concluding that the proponent has not established foundational reliability; or (3) they may admit the evidence with limitations, as some courts have done.²¹ While trial court determinations of admissibility are reviewed for an "abuse of discretion,"²² a misapprehension as to the applicable legal standard under Rule 702 is considered an abuse of discretion.²³

One issue facing judges when applying Rule 702's reliability requirements to feature comparison disciplines is whether such disciplines are "scientific," as compared to "technical" or "specialized." On the one hand, feature and pattern recognition experts distinguish themselves from lay persons on their ability to identify individuals based on the experts' knowledge, training, and expertise. Although lay people may be able to identify a known individual's handwriting²⁴ or may be able to distinguish between two clearly dissimilar fingerprints, courts have consistently ruled that feature and pattern recognition by trained individuals is, in fact, specialized knowledge.²⁵ On the other hand, some commentators have claimed that examiners wielding relatively more subjective experience- and judgment-based

Jonathan J. Koehler, *Forensics or Fauxrensic? Ascertaining Accuracy in the Forensic Sciences*, 49 Ariz. St. L. J. 1369 (2017) (explaining that, in most areas of forensic science, the necessary proficiency testing has not been done). Qualifications, proficiency, and reliability as applied are not entirely distinct concepts, and there is some overlap among them.

21. See, e.g., *United States v. Tibbs*, No. 2016-CF1-19431, 2019 WL 4359486 (D.C. Super. Ct. Sept. 5, 2019) (precluded the government from eliciting testimony identifying the recovered firearm as the source of the recovered cartridge and limiting expert's testimony to a conclusion that the firearm cannot be excluded as the source, based on the consistency of the class characteristics and microscopic toolmarks); *United States v. Adams*, 444 F. Supp. 3d 1248 (D. Or. 2020); *United States v. Shipp*, 422 F. Supp. 3d 762, 778–79 (E.D.N.Y. 2019) (disallowing a conclusion of a "match" with firearm evidence); *United States v. Rutherford*, 104 F. Supp. 2d 1190, 1193 (D. Neb. 2000) (disallowing the conclusion of a "match" of handwriting samples); *United States v. Van Wyk*, 83 F. Supp. 2d 515 (D.N.J. 2000); *United States v. Hines*, 55 F. Supp. 2d 62 (D. Mass. 1999) (same).

22. See *Gen. Elec. Co. v. Joiner*, 522 U.S. 136, 141–43 (1997).

23. See, e.g., *Koon v. United States*, 518 U.S. 81, 100 (1996) (noting that a trial court necessarily abuses its discretion when it fails to apply the correct legal standard or makes an error of law); *Jeff D. v. Otter*, 643 F.3d 278 (9th Cir. 2011) (same).

24. See, e.g., *Fed. R. Evid.* 901(b)(2) (permitting "nonexpert's opinion that handwriting is genuine, based on a familiarity with it that was not acquired for the current litigation").

25. See, e.g., *Shipp*, 422 F. Supp. 3d at 783 (firearm comparison); *United States v. Llera Plaza*, 188 F. Supp. 2d 549, 563–64 (E.D. Pa. 2002) (fingerprint comparison); *United States v. Mallory*, 902 F.3d 584, 593 (6th Cir. 2018) (handwriting comparison).

feature and pattern identification techniques are not “scientists” practicing “science.”²⁶

Of course, a field that is “technical” rather than “scientific” is still subject to the reliability requirements of Rule 702, as the Supreme Court held in *Kumho Tire Co. v. Carmichael*.²⁷ But the Supreme Court has not squarely addressed the types of factors a trial judge should consider in determining the foundational reliability of a nonscientific technical method.²⁸ The National District Attorneys Association suggested in its response to the 2016 PCAST Report that validation studies are not required for nonscientific, experience-based methods.²⁹ But as the PCAST committee responded, it is not clear what the alternative means of testing validity of such a method would be.³⁰ While some have argued that validity can be established through so-called “blind” verification by a second examiner or interlaboratory comparison, such comparisons would presumably focus mostly on repeatability, not accuracy. Thus, courts determining the foundational reliability of a feature comparison discipline method deemed nonscientific will have to determine how—if not through validation studies showing the limits of the method’s ability to get the right answer—the proponent of the method can establish its accuracy for its stated purpose by a preponderance of the evidence.

Reliability as applied. Once the proponent has demonstrated that the method is foundationally reliable (i.e., the method produces accurate, repeatable, and reproducible results on materials like those in the instant case), and that the examiner is qualified to conduct the method, the proponent must introduce additional evidence that the examiner in fact followed the method in the case at hand. The 2023 amendments to Rule 702 reaffirmed that reliability as applied, like other admissibility requirements, goes to admissibility and not merely weight and must be proven by a preponderance of the evidence.³¹ To prove that

26. See Letter from Nat’l Dist. Att’y’s Ass’n to President Obama in response to 2016 PCAST Report (Nov. 16, 2016), <https://perma.cc/5JX7-BU5V>. See generally discussion in section titled “Recurring Concerns with Forensic Feature Comparison Evidence” below.

27. *Kumho Tire Co. v. Carmichael*, 526 U.S. 137, 142 (1999).

28. See generally discussion in section titled “Prior Government and NGO Review of Forensic Feature Comparison Evidence” below (discussing critiques of PCAST and arguments about how and whether to calculate error rates for feature comparison disciplines).

29. See *id.* (noting the NDAA’s response to the 2016 PCAST Report).

30. See President’s Council of Advisors on Sci. & Tech. (PCAST), Exec. Office of the President, An Addendum to the PCAST Report on Forensic Science in Criminal Courts 1, 3 (Jan. 6, 2017), <https://perma.cc/TFP8-YUYS> [hereinafter PCAST Addendum] (Noting that some critics of PCAST “suggested that the validity and reliability of such a method could be established without actually empirically testing the method in an appropriate setting. Notably, however, none of these respondents identified any alternative approach that could establish the validity and reliability of a subjective forensic feature-comparison method.”).

31. Fed. R. Evid. 702 advisory committee’s note to 2023 amendment (“[M]any courts have held that the critical questions of the . . . application of the expert’s methodology, are questions of weight and not admissibility. These rulings are an incorrect application of Rules 702 and 104(a).”).

a laboratory or examiner reliably applied a method, the proponent might rely on various types of documentation, from internal validation studies (showing that the method works on the laboratory's equipment with the specific protocols, practices, and personnel of the specific forensic service provider) to case file documentation to examiner bench notes, reports, and testimony showing the conditions of the case and what the examiner did in the particular case.

This “reliability as applied” showing is different from foundational reliability or expert qualification. For example, a trial court might deem fingerprint comparison a foundationally reliable method and deem a particular examiner to be qualified based on experience, credentials, and proficiency tests. But if the comparison in the case is between a reference print and a particularly smudged or incomplete print that goes beyond the bounds of the method's empirically tested accuracy, the method might not be reliably applied. As the advisory committee note to the 2023 amendments emphasized, expert opinions must “stay within the bounds of what can be concluded from a reliable application of the expert's basis and methodology.”³² Another way that an opinion can exceed the bounds of a valid application of a method is if it relates to a different type of conclusion—for example, whether a note was made to look like a forgery, rather than whether two notes have the same author.³³ Still another way an opinion might be unreliable as applied is if the examiner claims certainty in their opinion, a concern the advisory committee considers “particularly pertinent to the testimony of forensic experts.” Instead, in deciding whether to admit the evidence, trial courts should “(where possible) receive an estimate of the known or potential rate of error of the methodology employed, based (where appropriate) on studies that reflect how often the method produces accurate results.”³⁴

Prior Government and NGO Review of Forensic Feature Comparison Evidence

This subsection offers an overview of governmental and nongovernmental organization (NGO) review of forensic feature comparison evidence over the past fifteen years. Judges should be aware of these reports both because they are frequently cited and because they can be helpful guides to areas of continuing research and controversy.

32. *See id.*

33. *See, e.g.,* *Almeciga v. Ctr. for Investigative Reporting, Inc.*, 185 F. Supp. 3d 401, 424 (S.D.N.Y. 2016) (noting that method for one purpose was not validated for other purpose and that a trial judge “should not [admit questioned document testimony] without carefully evaluating whether the examiner has actual expertise in regard to the specific task at hand”).

34. *See* Fed. R. Evid. 702 advisory committee's note to 2023 amendment.

The 2009 National Research Council Report. The push since the mid-2000s to improve feature comparison disciplines has resulted in part from the DNA exoneration movement, which exposed erroneous convictions based on embellished or misleading forensic evidence.³⁵ Additionally, the misidentification of Brandon Mayfield by the FBI as the perpetrator of the 2004 Madrid train bombings based on latent fingerprint analysis further spurred concern over non-DNA methods.³⁶ The National Research Council of the National Academies (NRC) began a formal review in 2006 of the state of non-DNA forensic feature comparison disciplines.³⁷ The result was a lengthy report published in 2009, *Strengthening Forensic Science in the United States: A Path Forward*.³⁸

The primary conclusion of the 2009 NRC Report was that, “[w]ith the exception of nuclear DNA analysis . . . no forensic [feature identification] method has been rigorously shown to have the capacity to consistently, and with a high degree of certainty, demonstrate a connection between evidence and a specific individual or source.”³⁹ Specifically, the NRC pointed out the relative subjectivity and nonreproducibility in non-DNA methods that are based primarily on examiner experience and judgment; the lack of validation studies to assess accuracy of these relatively subjective methods;⁴⁰ the lack of “population studies” that show the “variability” and rarity of a feature within a relevant population (critical to assessing the probative value of a “match” between two sets of observed features, like whorls and loops in a thumbprint or a certain shoe sole pattern);⁴¹ the lack of corrective measures to avoid examiner contextual bias;⁴² the “absence of a feedback mechanism” to alert examiners that a method

35. See generally Restoring Freedom, Innocence Project, <https://perma.cc/LMU4-87JU> (last visited Nov. 18, 2024) (explaining the categories of reasons underlying convictions of factually innocent defendants).

36. See Robert B. Stacey, Report on the Erroneous Fingerprint Individualization in the Madrid Train Bombing Case, Fed. Bureau of Investigation (FBI) (Jan. 2005), <https://perma.cc/JQ9W-RX7K>.

37. See generally Erin Murphy, What ‘Strengthening Forensic Science’ Today Means for Tomorrow: DNA Exceptionalism and the 2009 NAS Report, 9 L. Probability & Risk 7 (2010), <https://doi.org/10.1093/lpr/mgp030> (explaining the story of how DNA exonerations led to the Senate’s charge, and the forensic community’s request, to the NRC to write a report on non-DNA evidence); Nat’l Rsch. Council, Nat’l Acads., *Strengthening Forensic Science in the United States: A Path Forward* 1 (2009), <https://doi.org/10.17226/12589> [hereinafter 2009 NRC Report] (explaining Congress’s charge to them). We refer to the 2009 report as the “2009 NRC Report” to distinguish it from later reports from the National Academy of Sciences, even though some authors refer to the 2009 report as the “NAS Report.”

38. See 2009 NRC Report, *supra* note 37.

39. *Id.* at 7.

40. *Id.* at 8.

41. *Id.* at 149. See also *id.* at 154 (same concern for firearms and toolmarks), 163 (same for fiber evidence), 165 (handwriting), 184 (same concern for friction ridge patterns).

42. *Id.* at 8, 24.

might have produced a false positive or negative;⁴³ and the lack of “quantifiable measures of uncertainty” (i.e., an error rate).⁴⁴

Responses to the 2009 NRC Report varied, although the scientific community appeared largely supportive. Some pushed back against the report’s calling established methods into question.⁴⁵ On the other hand, organizations such as the American Academy of Forensic Sciences (AAFS) published a formal response “support[ing] the recommendations” of the report,⁴⁶ and other scientists followed suit.⁴⁷

2013–17: The National Commission on Forensic Science (NCFS). The 2009 NRC Report led to a broad debate over the validity and reliability of forensic science methods, and to the creation of the NCFS, a joint project of the Department of Justice (DOJ) and NIST. The NCFS’s thirty-seven members included forensic examiners, psychologists and other scientists, prosecutors, defense attorneys, academics, and judges.⁴⁸ Between 2014 and 2017, the NCFS approved twenty “recommendations” and twenty-three “views documents,” all available on the Commission’s website⁴⁹ and including recommendations requiring proficiency testing of examiners even in fields lacking an accreditation body; views on what federal prosecutors should disclose in pretrial discovery about forensic methods, above and beyond what is required by Federal Rule of Criminal Procedure 16;⁵⁰ and the view that forensic examiners should base their opinions only on

43. *Id.* at 149.

44. *Id.* at 23.

45. See generally Paul C. Giannelli, *The 2009 NAS Forensic Science Report: A Literature Review*, 48 *Crim. L. Bulletin* 378, 382, 385 (2012) (discussing responses to the 2009 NRC Report, including a prosecutor in front of the Senate Judiciary Committee insisting that the report was an “agenda-driven attack upon well-founded investigative techniques” and then-Senator Jeff Sessions’ comment that “I don’t think we should suggest that those proven scientific principles that we’ve been using for decades are somehow uncertain”).

46. See Am. Acad. of Forensic Scis. (AAFS), Response to the National Academy of Sciences’ “Forensic Needs” Report (Sept. 4, 2009), <https://perma.cc/K7B4-TWCP>.

47. See, e.g., Karen Kafadar, *Statistical Issues in Assessing Forensic Evidence*, 83 *Int’l Statistical Rev.* 111, 120 (2015), <https://doi.org/10.1111/insr.12069> (Expressing support for the 2009 NRC Report and opining that “[t]he scientific method is very general, and the principles of science apply to all branches, irrespective of the specific application (e.g. physics, chemistry or biology). If the scientific method has not been applied, then scientists from no branch can defend the results.”).

48. See U.S. Dep’t of Justice, National Commission on Forensic Science: Members (June 1, 2017), <https://perma.cc/V98X-BMQB> (listing members and biographies).

49. See U.S. Dep’t of Justice, National Commission on Forensic Science: Work Products Adopted by the Commission (July 23, 2018), <https://perma.cc/5XVZ-KN4F> (links to documents).

50. See *id.* These recommendations were written before the 2022 amendments to Rule 16, which (among other things) require parties to disclose not only written summaries of their expert’s anticipated testimony but “a complete statement” of the witness’s opinions, the bases and reasons for those opinions, the witness’s qualifications, and a list of other cases in which the witness had testified in the previous four years. Fed. R. Crim. P. 16 advisory committee’s note to 2022 amendment.

“task-relevant” information.⁵¹ While the NCFS’s charter expired in April 2017 and was not renewed by then-Attorney General Jeff Sessions,⁵² the views documents remain potentially persuasive documents wielded by parties litigating forensic evidence issues.

The 2016 PCAST Report. While the NCFS was still completing its work, PCAST issued a report in 2016 on forensic feature comparison evidence.⁵³ The report’s working group largely consisted of judges and academics from the “derivative sciences” (i.e., sciences like chemistry, statistics, genetics, and computer science underlying forensic methods), although the group sought input from “the Federal Bureau of Investigation (FBI) Laboratory and individual scientists at NIST, as well as from many other forensic scientists and practitioners, judges, prosecutors, defense attorneys, academic researchers, criminal-justice-reform advocates, and representatives of Federal agencies.”⁵⁴

The 2016 PCAST Report echoed the concerns of the 2009 NRC Report and focused on explaining what was missing to properly validate non-DNA forensic feature comparison methods. Specifically, PCAST explained that the only way to meaningfully test the validity of a method based primarily on examiner judgment and experience (where the method’s steps are largely an inscrutable “black box” to those on the outside) is through “black box” studies where the method’s accuracy is tested on samples where the tester knows the right answer.⁵⁵ In turn, PCAST explained that any statement about a method’s reliability, or the certainty underlying an examiner’s conclusion, should not overstate the probative value of the evidence beyond what is “demonstrated by empirical evidence” based on error-rate studies.⁵⁶ “For example,” PCAST wrote, “if the false positive rate of a method has been found to be 1 in 50, experts should not imply that the method is able to produce results at a higher accuracy.”⁵⁷

The PCAST Report then covered what a well-designed error rate study requires: samples and populations representative of real casework; a sufficiently

51. See U.S. Dep’t of Justice, Views of the Commission Ensuring That Forensic Analysis Is Based Upon Task-Relevant Information (Dec. 8, 2015), <https://perma.cc/U2JF-QSCP>. This document was heavily influenced by, and frequently cited, the work of psychologist Itiel Dror (who spoke to the commission), who has written on contextual bias in forensic feature comparison methods such as DNA mixture interpretation and latent fingerprint analysis. See Itiel E. Dror & Greg Hampikian, *Subjectivity and Bias in Forensic DNA Mixture Interpretation*, 51 Sci. & Justice 204 (2011), <https://doi.org/10.1016/j.scijus.2011.08.004>, and Itiel E. Dror et al., *Cognitive Issues in Fingerprint Analysis: Inter- and Intra-Expert Consistency and the Effect of a ‘Target’ Comparison*, 208 Forensic Sci. Int’l 10 (2011), <https://doi.org/10.1016/j.forsciint.2010.10.013>.

52. See U.S. Dep’t of Justice, National Commission on Forensic Science, <https://perma.cc/V6GT-YSWN> (noting charter expiration in April 2017).

53. See 2016 PCAST Report, *supra* note 20.

54. *Id.* at 2.

55. *Id.* at 5.

56. *Id.* at 6.

57. *Id.*

large sample size; “blind” testing; and no mid-stream change in protocol. In turn, PCAST opined on the proper way to calculate a method’s error rate. First, it insisted that the false positive rate be calculated by determining the number of false inclusions among an examiner’s *conclusive* examinations (exclusions or declarations of a “match”) rather than all examinations (including “inconclusive” determinations).⁵⁸ Second, the reported error rate should be the upper bound of a “confidence interval” around the study error rate to account for measurement uncertainty, so that the factfinder has a sense of how high the false positive rate might be.⁵⁹ Using these criteria, PCAST ultimately found insufficient evidence of foundational validity for several feature comparison disciplines—toolmark analysis, handwriting analysis, bitemark analysis, footwear analysis, microscopic hair analysis, and deconvolution of DNA mixtures over three contributors.⁶⁰ The only non-DNA feature comparison method PCAST deemed foundationally valid was latent print analysis, but added several caveats.⁶¹

The 2016 PCAST Report garnered formal responses from numerous organizations, some supportive and some critical.⁶² PCAST issued an addendum to its report in January 2017 responding to critiques.⁶³ After the addendum, additional critiques⁶⁴ and defenses⁶⁵ of PCAST continued. A 2018 article by Professor Jay Koehler, directed at the judiciary, offers a helpful and measured summary of the critiques of PCAST, possible responses, lingering debates, and issues for

58. *Id.* at 51.

59. *Id.*

60. *Id.* at 7.

61. *Id.* at 9–10.

62. See, e.g., Nat’l Dist. Att’y’s Ass’n, *supra* note 26 (arguing that not all methods are “scientific,” and not all methods need to be supported by validation studies); Fed. Bureau of Investigation (FBI), Comments on President’s Council of Advisors on Science and Technology REPORT TO THE PRESIDENT Forensic Science in Federal Criminal Courts: Ensuring Scientific Validity of Pattern Comparison Methods [PCAST Report] (Sept. 20, 2016), <https://perma.cc/B5S9-XSG3> (criticizing PCAST for not taking into account studies that the FBI argued met the criteria for appropriate black box studies).

63. See PCAST Addendum, *supra* note 30.

64. See, e.g., Ted Robert Hunt, *Scientific Validity and Error Rates: A Short Response to the PCAST Report*, 86 Fordham L. Rev. Online 24, 26 (2018) (agreeing that methods must be tested empirically but deeming PCAST’s views of an appropriate test as too narrow); U.S. Dep’t of Justice, *Justice Department Publishes Statement on 2016 President’s Council of Advisors on Science and Technology Report* (Jan. 13, 2021), <https://perma.cc/8J6L-L4DN> (arguing that black box studies are not always needed and that PCAST’s criteria are too narrow).

65. See Letter from Democracy Forward Foundation on behalf of Union of Concerned Scientists, requesting correction under the Information Quality Act (regarding U.S. Dep’t of Justice’s statement on the PCAST Report) (June 24, 2021), <https://perma.cc/MQD8-K835> (defending PCAST and demanding that DOJ retract its statement); Innocence Project Staff, *Innocence Project Calls on Department of Justice to Retract Statement on PCAST Report* (Feb. 19, 2021), <https://perma.cc/SS23-4K5F> (same).

judges to resolve in light of these debates.⁶⁶ For their part, statisticians and other scientists have likewise urged the forensic community to follow PCAST's recommendations.⁶⁷

Notwithstanding its continued criticism of PCAST, DOJ has voluntarily made several efforts to improve its forensic practices in response to PCAST and other reports. From 2017 onward, the department has implemented several new initiatives, including an updated code of professional responsibility, publication of research needs, testimony monitoring, and creation of a laboratory needs working group.⁶⁸ FBI representatives have also reported such DOJ initiatives being implemented to strengthen the accuracy of FBI experts' testimony.⁶⁹ DOJ does continue to use and endorse the term "source identification," notwithstanding concerns expressed by others to the Advisory Committee on the Rules of Evidence that this term is synonymous with a claim of individualization, which DOJ reports its experts no longer testify to in court.⁷⁰

The forensic community has responded to PCAST in part by promulgating standards to improve feature comparison disciplines through a new entity, NIST's Organization of Scientific Area Committees (OSAC). OSAC is composed of hundreds of forensic examiners along with lawyers, judges, psychologists, statisticians, and laboratory directors. Created in 2014, its mission is "to strengthen the nation's use of forensic science by facilitating the development

66. Jonathan Jay Koehler, *How Trial Judges Should Think About Forensic Science Evidence*, 102 *Judicature* 29 (2018), <https://perma.cc/DLM4-CXCL>.

67. See, e.g., Suzanne Bell et al., *A Call for More Science in Forensic Science*, 115 *Proc. Nat'l Acad. Scis.* 4541, 4541 (2018), <https://doi.org/10.1073/pnas.1712161115> (Discussing the 2009 NRC and PCAST reports with approval and opining that "[f]orensic science is at a crossroads. It is torn between the practices of science, which require empirical demonstration of the validity and accuracy of methods, and the practices of law, which accept methods based on historical precedent even if they have never been subjected to meaningful empirical validation."); Thomas D. Albright, *The US Department of Justice Stumbles on Visual Perception*, 118 *Proc. Nat'l Acad. Scis.* e2102702118 (2021), <https://doi.org/10.1073/pnas.2102702118> (offering a neuroscientist's critique of DOJ's refusal to adopt PCAST's findings).

68. See U.S. Dep't of Justice, *Forensic Science: Publications*, <https://perma.cc/BLR3-WNTK> (last visited Nov. 20, 2024) (listing priorities, publications, and initiatives).

69. See, e.g., Alice R. Isenberg & Cary T. Oien, *Scientific Excellence in the Forensic Science Community*, 86 *Fordham L. Rev. Online* 39 (2018) (discussing testimony monitoring as an example of a recent initiative, as well as the FBI's existing accreditation requirements); Andrew D. Goldsmith, *The Reliability of the Adversarial System to Assess the Scientific Validity of Forensic Evidence*, 86 *Fordham L. Rev. Online* 16 (2018) (noting that the FBI has eliminated use of the terms "reasonable degree of scientific certainty" and similar statements; eliminated claims of zero error; prohibited examiners from citing the number of examinations conducted as an indication of the accuracy of their conclusion; and created a testimony monitoring program). However, DOJ's policies apply only to its own witnesses. If witnesses from local or state agencies testify in a federal case, they may not adhere to the DOJ policy.

70. See Judicial Conference Advisory Committee on the Federal Rules of Evidence, *Minutes of the Meeting of May 3, 2019*, 1, 20–23 (May 3, 2019), <https://perma.cc/2QRN-7LQ8> (disagreement on the FBI's continued use of this term and whether it is different from individualization).

and promoting the use of high-quality, technically sound standards.”⁷¹ OSAC both creates new standards and reviews standards set by other standards development organizations (SDOs), such as the American Academy of Forensic Science’s (AAFS) Academy Standards Board (ASB),⁷² ASTM International (formerly the American Society of Testing and Materials),⁷³ and the National Fire Protection Association (NFPA).⁷⁴ Standards that have sufficient “technical merit” are placed on OSAC’s “registry.”⁷⁵ Judges should be aware of OSAC’s activities and output, as well as controversy over the quality of its registry standards, given that a standard’s inclusion on the registry may be touted by one side or another as proof that a discipline is governed by standards, one of the “*Daubert* factors.”⁷⁶

In addition to OSAC, other governmental and nongovernmental organizations independent of the forensic examiner or legal communities⁷⁷ continue efforts to improve forensic science. The Center for Statistics and Applications in

71. NIST’s Organization of Scientific Area Committees for Forensic Science (OSAC), *About Us*, Nat’l Inst. of Standards & Tech. (NIST), U.S. Dep’t of Commerce, <https://perma.cc/9NJ8-7L2F> (last visited Nov. 20, 2024).

72. See Academy Standards Board, Am. Acad. of Forensic Sci. (AAFS), <https://perma.cc/U74R-Y4SM> (last visited Nov. 20, 2024).

73. See ASTM International, <https://perma.cc/25VC-ZCQK> (last visited Nov. 20, 2024).

74. See Nat’l Fire Prot. Ass’n (NFPA), NFPA Codes and Standards, <https://perma.cc/G24D-JGNZ> (last visited Nov. 20, 2024).

75. See NIST’s OSAC, *About Us*, *supra* note 71 (“OSAC also reviews standards and posts high quality ones to the OSAC Registry [<https://perma.cc/L2SZ-L7X3> (last visited Nov. 20, 2024)]. Inclusion on this registry indicates that a standard is technically sound and that laboratories should consider adopting them.”).

76. Three OSAC scientists published commentary criticizing OSAC for promulgating “vacuous” standards that merely require laboratories to *have* standards, rather than offering meaningful guidance on what the standards should be. These critics claimed that placing such standards on the OSAC registry allows disciplines to misleadingly claim they are governed by standards. Geoffrey Stewart Morrison, Cedric Neumann & Patrick Henry Geoghegan, *Vacuous Standards—Subversion of the OSAC Standards-Development Process*, 2 Forensic Sci. Int’l: Synergy 206 (2020), <https://doi.org/10.1016/j.fsisyn.2020.06.005>. In response, several members of ASB argued that the standards were an improvement, had to go through several rounds of comments from task groups and the public, and can be revised. See Linton A. Mohammed et al., *Response to Vacuous Standards Subversion of the OSAC Standards Development Process*, 3 Forensic Sci. Int’l: Synergy 100145 (2021), <https://doi.org/10.1016/j.fsisyn.2021.100145>. A reply to the response, signed by eleven scientists, including several OSAC members, argued that the issue is the content of the standards, not the process that led to them, and ended with an admonition to judges to “not accept at face value claims of scientific validity based on the fact that published standards have been followed. We would encourage courts to enquire further so as to ascertain whether those standards are fit for purpose.” See Geoffrey Stewart Morrison et al., *Reply to Response to Vacuous Standards—Subversion of the OSAC Standards-Development Process*, 3 Forensic Sci. Int’l: Synergy 100149 (2021), <https://doi.org/10.1016/j.fsisyn.2021.100149>.

77. There are also professional organizations within the forensic examiner and legal communities that work to improve and critique forensic techniques, such as the Forensic Justice Project, <https://perma.cc/RS96-DK5U>, and the American Academy of Forensic Sciences, <https://perma.cc/CSF3-69L4>.

Forensic Evidence (CSAFE), a federally funded research and training center that is the only one of three NIST Centers of Excellence dedicated to building the scientific and statistical foundations of feature and pattern comparison evidence, remains active in nationwide efforts to review and improve forensic feature comparison methods. In particular, many of its members are statisticians developing more objective means of rendering feature and pattern comparison based on automated methods.⁷⁸ Likewise, the American Association for the Advancement of Science (AAAS) holds conferences and publishes articles related to forensic science issues.⁷⁹

Recurring Concerns with Forensic Feature Comparison Evidence

Before this reference guide delves into specific disciplines, this section flags several recurring issues with respect to many non-DNA feature comparison methods that judges will face.

Terminology and Testimony Issues

Overclaims of certainty and challenges to claims of “matches” or source attribution. Challenges to non-DNA feature comparison expert testimony sometimes relate to the terms used by experts to describe their conclusions. One issue is claims of certainty. The advisory committee’s notes to the 2023 amendments to Rule 702 state that “[f]orensic experts should avoid assertions of absolute or one hundred percent certainty.”⁸⁰ The 2009 NRC Report and 2016 PCAST Report also urged against such statements,⁸¹ and DOJ’s guidelines prohibit them.

Another frequently challenged word in feature comparison reports and testimony is “match.” When an expert deems two sets of features to be consistent

78. See generally Center for Statistics and Applications in Forensic Evidence (CSAFE), <https://perma.cc/UW29-6XDB> (linking to studies and ongoing research).

79. For example, the AAFS Center for Scientific Responsibility and Justice held a conference on forensics in 2019 on the tenth anniversary of the 2009 NRC Report. See Am. Ass’n for the Advancement of Sci. (AAAS), An Update on Strengthening Forensic Science in the United States: A Decade of Development [Forensic Conference 2019], <https://perma.cc/8NXY-9WHK>. See generally Am. Ass’n for the Advancement of Sci. (AAAS), “Forensic” Search Results, <https://perma.cc/9ZRZ-4RQH> (listing publications and events).

80. Fed. R. Evid. 702 advisory committee’s note to 2023 amendment.

81. See, e.g., 2009 NRC Report, *supra* note 37, at 47–48 (noting testimony of some experts as to “unfailing certainty” and describing “failure to acknowledge uncertainty” as a problem in feature comparison methods); PCAST Report, *supra* note 20 at 6 (criticizing experts’ “[s]tatements claiming or implying greater certainty than demonstrated by empirical evidence”).

with each other, they might testify that the evidence and reference samples “match.” But the term “match” might inaccurately suggest both an air of certainty in the expert’s determination of the consistency features, and that the mere consistency of a set of compared features means that the two items necessarily share a common source. For example, if two fingerprints “match” at six points, they might not “match” if the examiner looked at ten more points. For a discipline like DNA, with robust population data and quantified “match” statistics, examiners do not need to fall back on terms like “match” or “identification”; they can just present a statistic and allow the factfinder to make their own determination of whether a suspect is the source of the DNA. But for more subjective methods without such data, an examiner’s opinion that two patterns “match” or share the same source is more fraught.

For similar reasons, government and NGO reports have also voiced concern over use of other terms that imply that two patterns definitely have the same source, such as “individualization,” “similar in all respects tested,” or “to the exclusion of all others,” given their “profound effect on how the trier of fact in a criminal or civil matter perceives and evaluates scientific evidence.”⁸² For example, the 2009 NRC Report expressed concern that imprecise terminology in microscopic hair analysis, such as “associated with,” could imply individualization and a “match” without necessary confirmation from mitochondrial DNA (mtDNA) analysis.⁸³ Indeed, an internal FBI study found that “of 80 hair comparisons that were ‘associated’ through microscopic examinations, 9 of them (12.5 percent) were found in fact to come from different sources when reexamined through mtDNA analysis.”⁸⁴ The 2016 PCAST Report likewise urged that “quantitative information about the reliability of methods be stated clearly in expert testimony,” and ambiguous language such as “match” and “identification” be avoided, as this could be misinterpreted as scientifically supported individualization.⁸⁵ PCAST also discouraged examiners from testifying as to the “uniqueness” of features rather than focusing on the extent to which evidence of consistency in features supports an inference that the features or patterns share a common source.⁸⁶

Other scientific organizations, such as the American Statistical Association (ASA), have voiced similar warnings about language suggesting identification or source attribution.⁸⁷ To avoid factfinders making inaccurate inferences about

82. 2009 NRC Report, *supra* note 37, at 21. *See also id.* at 7, 43 (“individualization”); 176 (“exclusion of all others”).

83. *Id.* at 161.

84. *Id.* at 160–61 (discussing Max Houck & Bruce Budowle, *Correlation of Microscopic and Mitochondrial DNA Hair Comparisons*, 47 J. Forensic Scis. 964 (2002)).

85. 2016 PCAST Report, *supra* note 20, at 121–22.

86. *Id.* at 61–62.

87. *See* Am. Stat. Ass’n, American Statistical Association Position on Statistical Statements for Forensic Evidence 1, 4–5 (Jan. 2, 2019), <https://perma.cc/GHH3-VY4G> (“The ASA strongly discourages statements to the effect that a specific individual or object is the source of the forensic

certainty, the ASA further suggests requiring experts to make clear in their testimony that “it is possible that other individuals or objects may possess or have left a similar set of observed features” and to acknowledge “both in testimony and in written reports” a method’s “absence of models and empirical evidence.”⁸⁸ While some testimony guides direct examiners to limit their claims to stating that evidence is “strong support” for identification (as compared to a statement of categorical source attribution), such statements themselves should be backed up by empirical data about the rarity of features such that the examiner’s observation of them justifies such language in a given case.

Still other terms have been critiqued as simply inappropriate. For example, the NCFS published a consensus views document that terms like “reasonable degree of scientific certainty” and “reasonable degree of ballistic certainty” have no scientific basis and should not be used by legal actors or examiners.⁸⁹ The NCFS also published a consensus views document on “Inconsistent Terminology,” which discussed inconsistent and misleading uses across disciplines of words like “perimortem,” “inconclusive,” or implicitly overclaiming phrases like “there could be another individual somewhere in the world” with the same feature or pattern.⁹⁰ In federal court, some examiners will be limited by DOJ’s guidelines (relatively new as of this writing) on uniform language for testimony and reports, which typically require examiners to report a source identification, a source exclusion, or an inconclusive determination.⁹¹ Some commentators urge a retreat from such categorical language, given the lack of data about the rarity of features in the population that would presumably be needed to scientifically support any claim of individualization. Instead, some commentators urge examiners to only use scientifically supported statements about the strength of the evidence, in the form of likelihood ratios or other statements that take into account population data showing the rarity of features in the relevant population.⁹²

science evidence. Instead, the ASA recommends that reports and testimony make clear that, even in circumstances involving extremely strong statistical evidence, it is possible that other individuals or objects may possess or have left a similar set of observed features. . . . The ASA recommends that the absence of models and empirical evidence be acknowledged both in testimony and in written reports.”).

88. *Id.*

89. See U.S. Dep’t of Justice, Nat’l Comm’n on Forensic Sci. (NCFS), Work Products Adopted by the Commission (July 23, 2018), <https://perma.cc/64DR-4VDA> (links to documents).

90. See U.S. Dep’t of Justice, NCFS, Final Draft Views on Inconsistent Terminology, <https://perma.cc/9CT6-VY3U> (last visited Nov. 26, 2024).

91. See, e.g., U.S. Dep’t of Justice, Uniform Language for Testimony and Reports for the Forensic Firearms/Toolmarks Discipline Pattern Examination (2023), <https://perma.cc/4N57-TDAG>.

92. See, e.g., Geoffrey Stewart Morrison et al., *Calculation of Likelihood Ratios for Inference of Biological Sex from Human Skeletal Remains*, 3 Forensic Sci. Int’l: Synergy 100202 (2021), <https://doi.org/10.1016/j.fsisy.2021.100202> (urging use of likelihood ratios (LRs) instead of statements of

Conflating related but distinct scientific principles such as “reliability” and “validity.” Although courts often use the terms *validity* and *reliability* interchangeably, the terms have distinct meanings in different scientific disciplines. Validity typically refers to the ability of a test to “accurately measure[] what it is intended to measure.”⁹³ Reliability, in scientific and statistical terms, refers to the extent to which a method “produces largely consistent results when properly applied.”⁹⁴ Reliability can be further broken down into repeatability (“intra-examiner reliability,” or the extent to which the same examiner reaches the same results under the same conditions) and reproducibility (“inter-examiner reliability,” or the extent to which other examiners reach the same result using the same method).⁹⁵ The Supreme Court acknowledged these distinctions in *Daubert*, but the Court indicated that it was using the term reliability in a different sense. The Court wrote that its concern was “evidentiary reliability—that is, trustworthiness. . . . In a case involving scientific evidence, *evidentiary reliability* will be based upon *scientific validity*.”⁹⁶ Thus, judges should be aware that when the 2016 PCAST Report refers to “foundational validity” and “validity as applied,” it is speaking to the same questions as *Daubert*. At the same time, other experts might speak of a method’s reliability solely in terms of its reproducibility and repeatability, rather than its accuracy in getting the right answer. Judges should be aware of the context in which the terms are used.

“Class” versus “subclass” versus “individual” characteristics. Many non-DNA forensic feature comparison techniques purport to distinguish among “class,” “subclass,” and “individual” characteristics. Class and sub-class characteristics are believed to be shared by a group of persons or objects (e.g., ABO blood types, or the tool-mark patterns seen on a particular kind of Glock handgun).⁹⁷ So-called individual characteristics are thought to be unique to an object or person, to the

source identification); David H. Kaye, *The Nikumaroro Bones: How Can Forensic Science Assist Fact-finders?*, 6 Va. J. Crim. L. 101 (2018) (same).

93. See Hal S. Stern, Maria Cuellar & David Kaye, *Reliability and Validity of Forensic Science Evidence*, 16 Significance 21, 22 (2019), <https://doi.org/10.1111/j.1740-9713.2019.01250.x>.

94. *Id.*

95. *Id.* at 22–23.

96. *Daubert v. Merrell Dow Pharms., Inc.*, 509 U.S. 579, 590 n.9 (1993) (“We note that scientists typically distinguish between ‘validity’ (does the principle support what it purports to show?) and ‘reliability’ (does application of the principle produce consistent results?) . . .”).

97. See Jan S. Bashinski & Joseph L. Peterson, *Forensic Sciences, in Local Government: Police Management* 556 n.74 (William Geller & Darrel Stephens eds., 4th ed. 2004):

The forensic scientist first investigates whether items possess similar ‘class’ characteristics—that is, whether they possess features shared by all objects or materials in a single class or category. (For firearms evidence, bullets of the same caliber, bearing rifling marks of the same number, width, and direction of twist, share class characteristics. They are consistent with being fired from the same *type* of weapon.) The forensic scientist then attempts to determine an item’s ‘individuality’—the features that make one thing different from all others similar to it, including those with similar class characteristics.

exclusion of all others. The term *match* is especially ambiguous with respect to these types of methods because a match could mean merely that the two patterns are thought to come from the same class or subclass, or it could mean that the two patterns are thought to come from the same source because they share individual characteristics. Expert opinions involving “individual,” “subclass,” and “class” characteristics raise different issues, including whether the determination rests on a firm scientific foundation⁹⁸ and how large a class or subclass is (for determining the probative value of a determination that two compared items are part of the same class or subclass).⁹⁹

Where an examiner seeks to testify to observing an individual characteristic, judges might further inquire into the assumptions underlying that description as part of a *Daubert* inquiry. The examiner is presumably assuming that a particular observed feature is unique, raising the need for empirical data suggesting uniqueness. The examiner’s use of that term also presumably reflects an assumption that a single “individual” characteristic is sufficient to make a “source identification” conclusion, given that no other item in the world has that one characteristic. Some disciplines might be able to sustain that assumption, while others might not.

Presenting conclusions in quantitative terms based solely on examinations conducted by the examiner. Forensic examiners from some disciplines and laboratories sometimes appeal to the sheer number of examinations they have done as a sufficient empirical basis for a source conclusion statement. For example, an examiner might claim that “in 10 years of practice they have examined 5,000 items like this and have never seen such a high degree of similarity,” or that “the probability of observing such a strong similarity if items do not have a common source is less than 1 in a 1,000.” The FBI, for its part, has recently affirmed that its testimony monitoring program seeks to eliminate such claims made by testifying experts.¹⁰⁰

The fallacy of the transposed conditional (the “prosecutor’s fallacy”). Forensic examiners, like all people, occasionally fall prey to statistical fallacies in drawing conclusions. One recurring example in the feature comparison context is the so-called fallacy of the transposed conditional. A conditional probability is the

98. See Michael Saks & Jonathan Koehler, *The Individualization Fallacy in Forensic Science Evidence*, 61 Vand. L. Rev. 199 (2008).

99. See Margaret A. Berger, *Procedural Paradigms for Applying the Daubert Test*, 78 Minn. L. Rev. 1345, 1356–57 (1994):

We allow eyewitnesses to testify that the person fleeing the scene wore a yellow jacket and permit proof that a defendant owned a yellow jacket without establishing the background rate of yellow jackets in the community. Jurors understand, however, that others than the accused own yellow jackets. When experts testify about samples matching in every respect, the jurors may be oblivious to the probability concerns if no background rate is offered, or may be unduly prejudiced or confused if the probability of a match is confused with the probability of guilt, or if a background rate is offered that does not have an adequate scientific foundation.

100. See Goldsmith, *supra* note 69 (noting the FBI’s position on this).

probability of one event A “given” another event B—for example, the probability that a playing card is a queen, given that it is a heart, is 1 in 13. If you “transposed” this conditional it would be the probability of B given A—for example, the probability a card is a heart, given that it is a queen, which is instead 1 in 4. Conflating the two as equal is the fallacy of the transposed conditional. In the forensic context, the fallacy would be to conflate one conditional probability—“the chance that he would match the crime scene evidence given that he is innocent,” with the transposed conditional—“the chance that he is innocent given that he matches the crime scene evidence.” Imagine, for example, that there was A positive blood found at a crime scene, and that the arrested suspect has A positive blood (along with 30% of the rest of the population). We might say that the chance a random person would “match” the blood evidence, assuming he is innocent, is 30%. But it’s definitely not true that the chance a person is innocent, given the mere fact that he “matches” the blood evidence, is 30%. In a population of a million people, a full 300,000 people would “match” the blood evidence simply by coincidence. The chance that any one of these 300,000 random people is innocent, given that they “match,” is certainly more than 30%; indeed, it is overwhelmingly high.

But when the probabilities become very small, people have a hard time with this concept. In the DNA context, the error would be as follows. Imagine a random match probability (say, 1 in a million), which is the probability that, given a person is not the source of the DNA, they will “match” by sheer chance. This means we would expect that every 1 in a million people would have this profile, and that, in a country of 300 million people, around 300 would “match” by chance. The fallacy would be to inaccurately present it as the probability that, given a person “matches” the DNA at a crime scene, there is only a 1 in a million chance they are not the source.¹⁰¹ In the shoe print context, the error would be similar. If one starts with the statement, “given a randomly selected person, the probability they will have these particular chevron and wave patterns on their shoe is small,” the fallacy would be to claim that “given this person had these particular chevron and wave patterns on their shoe, the chance that they are not the source is small.”

Reliability Concerns

Concerns about foundational reliability. As reflected in the governmental and NGO reports discussed in the previous subsection titled “Prior Government and NGO Review of Forensic Feature Comparison Evidence,” lingering concerns

101. See, e.g., *McDaniel v. Brown*, 558 U.S. 120 (2010) (noting that the prosecutor and the government’s DNA expert both engaged in the fallacy of the transposed conditional in their statements before the jury); Andrea Roth, *Safety in Numbers?: Deciding When DNA Alone Is Enough to Convict*, 85 N.Y.U. L. Rev. 1130 (2010) (explaining the fallacy in laypersons’ terms).

exist about the foundational reliability of some non-DNA forensic feature comparison methods. These are important concerns, as “it has become apparent, over the past decade, that faulty forensic feature comparison has led to numerous miscarriages of justice.”¹⁰² In brief, critics focus on four main issues.

First, non-DNA methods lack variability data that show how rare the compared features are in the relevant population. Thus, it is difficult to know merely from the presence of shared features between two items how strong the inference should be that the two items share a common source.¹⁰³ While some examiners might rely on their own casework experience to opine about the likely frequency of features in the relevant population, the 2016 PCAST Report cautioned against reliance on examiner experience on this front without empirical data.¹⁰⁴

Second, these methods are relatively subjective compared to DNA.¹⁰⁵ Because they are based largely on individual examiner judgment and experience, the feature comparisons themselves are sometimes critiqued as more prone to contextual bias and less able to be repeated and reproduced both by the same examiner and by other examiners, which are hallmarks of good science (reproducibility and repeatability).

Third, critics argue there are too few well-designed validation studies indicating the extent of a method’s accuracy. Put differently, these methods have not been sufficiently shown through well-designed empirical testing to consistently get the right answer. Regardless of how subjective or black box-like a method is, the method can still be tested against a known truth to determine how often the

102. 2016 PCAST Report, *supra* note 20, at 44 (citing exoneration compilations); see also Kori Khan & Alicia Carriquiry, *Hierarchical Bayesian Non-response Models for Error Rates in Forensic Black-Box Studies*, 381 *Phil. Transactions Royal Soc’y A: Mathematical, Physical, & Eng’g Scis.* 1, 2 (2023), <https://doi.org/10.1098/rsta.2022.0157> (explaining that a primary cause of these wrongful convictions is the “ad hoc, subjective methods to evaluate evidence, and the exaggerated claims by expert witnesses at trial”).

103. See, e.g., 2009 NRC Report, *supra* note 37, at 149 (critiquing lack of variability studies in non-DNA methods); Robin Mejia et al., *What Does a Match Mean? A Framework for Understanding Forensic Comparisons*, 16 *Significance* 25–28 (2019), <https://doi.org/10.1111/j.1740-9713.2019.01251.x> (noting the lack of population data to explain the rarity of characteristics and the resulting inability to accurately determine the probative value of a match).

104. 2016 PCAST Report, *supra* note 20, at 55 (“The frequency with which a particular pattern or set of features will be observed in different samples, which is an essential element in drawing conclusions, is not a matter of ‘judgment.’ It is an empirical matter for which only empirical evidence is relevant.”).

105. See, e.g., *id.* at 47 (“Objective methods are, in general, preferable to subjective methods. Analyses that depend on human judgment (rather than a quantitative measure of similarity) are obviously more susceptible to human error, bias, and performance variability across examiners.”). For more on the need for objective measures, see Karen Kafadar, *The Need for Objective Measures in Forensic Science*, 16 *Significance* 16 (2019), <https://doi.org/10.1111/j.1740-9713.2019.01249.x>. Of course, DNA analysis itself is also subjective in many respects. See generally Erin Murphy, *The Art in the Science of DNA: A Layperson’s Guide to the Subjectivity Inherent in Forensic DNA Typing*, 58 *Emory L.J.* 489 (2008).

method gets the right answer.¹⁰⁶ Such testing can be used to calculate an error rate, giving factfinders a better sense of how strongly the evidence supports the inference the examiner has drawn. The 2023 amendments to Rule 702 reflect the growing consensus that courts must be willing to “receive an estimate” of the error rates on methodologies employed. The advisory committee’s notes’ emphasis on error rates is consistent with the independent consensus positions of multiple national committees¹⁰⁷ as reflected in the 2009 NRC Report,¹⁰⁸ the 2016 PCAST Report,¹⁰⁹ and other recent reports from scientific review boards and experts.¹¹⁰

The 2016 PCAST Report in particular concluded that non-DNA feature comparison specialties vary widely in terms of well-designed validation studies. The report, after reviewing existing validation studies for a variety of disciplines, concluded that friction ridge analysis (fingerprint comparison) and analysis by expert systems of certain DNA mixtures have proof of foundational reliability with a known error rate,¹¹¹ while the validity of other methods (such as microscopic hair comparison, toolmark analysis, and bitemark analysis) had not been

106. See 2016 PCAST Report, *supra* note 20, at 5–6 (noting that “black box” methods can and must be tested).

107. See generally Paul C. Giannelli, *Forensic Science: Daubert’s Failure*, 68 Case W. Rsrv. L. Rev. 869 (2018) (discussing the conclusions of 2009 NRC Report, PCAST, and the NCFS).

108. See discussion at section titled “Prior Government and NGO Review of Forensic Feature Comparison Evidence” above.

109. See *id.*

110. See, e.g., Karen Kafadar, *Statistical Issues in Assessing Forensic Evidence*, 83 Int’l Stat. Rev. 111, 120 (2015), <https://doi.org/10.1111/insr.12069> (Expressing support for the 2009 NRC Report and opining that “[t]he scientific method is very general, and the principles of science apply to all branches, irrespective of the specific application (e.g. physics, chemistry or biology). If the scientific method has not been applied, then scientists from no branch can defend the results.”); William Thompson et al., Am. Ass’n for the Advancement of Sci. (AAAS), *Forensic Science Assessments: A Quality and Gap Analysis: Latent Fingerprint Examination 7–8*, 43–44 (2017), <https://www.aaas.org/report/latent-fingerprint-examination> [hereinafter AAAS Fingerprint Report] (noting the need to quantify uncertainty in the method); Melissa Taylor et al., NIST, *Forensic Handwriting Examination and Human Factors: Improving the Practice Through a Systems Approach* (2020), <https://doi.org/10.6028/NIST.IR.8282> [hereinafter 2020 NIST Report] (same); Kelly Sauerwein et al., *Bitemark Analysis: A NIST Scientific Foundation Review* (2023), <https://doi.org/10.6028/NIST.IR.8352> [hereinafter NIST Bitemark Report] (same).

111. 2016 PCAST Report, *supra* note 20, at 9 (“PCAST finds that latent-fingerprint analysis is a foundationally valid subjective methodology—albeit with a false positive rate that is substantial and is likely to be higher than expected by many jurors based on longstanding claims about the infallibility of fingerprint analysis.”). See also PCAST Addendum, *supra* note 30, at 4:

[T]he friction-ridge discipline . . . has set an excellent example by undertaking both (i) path-breaking black-box studies to establish the validity and degree of reliability of latent-fingerprint analysis, and (ii) insightful “white-box” studies that shed light on how latent-print analysts carry out their examinations, including forthrightly identifying problems and needs for improvement. PCAST also applauds ongoing efforts to transform latent-print analysis from a subjective method to a fully objective method.

supported by even one appropriately designed validation study.¹¹² More recent studies since the 2016 PCAST Report have echoed these concerns.¹¹³

Fourth, some examiners have used methods for a purpose or context beyond its scientific foundations. Even if a method might be foundationally valid for one modest purpose (such as determining whether two projectiles were fired from the same brand and model of firearm), it might not be foundationally valid for a more ambitious purpose (such as determining the likelihood that two projectiles were fired from the same firearm). Similarly, a handwriting comparison method might be foundationally valid for determining whether two exemplars were written by a different person (exclusion), but not for determining the likelihood that they were written by the same person. The more ambitious an examiner's claim on the witness stand, the more the concern that the method has been stretched beyond its scientific foundations. This could also be an issue with the method's reliability as applied, but it could be seen as an issue of foundational reliability if the method itself is simply not fit for the purpose it is used for.

These continuing issues with non-DNA feature comparison methods are likely in part a reflection of their origins outside the derivative sciences. For the most part, non-DNA feature comparison methods did not develop as “scientific” methods; rather, they grew from forensic practices that did not have governing

112. See PCAST Addendum, *supra* note 30, at 5 (“In its report, PCAST stated that it found no empirical studies whatsoever that establish the scientific validity or degree of reliability of bitmark analysis as currently practiced. To the contrary, it found considerable literature pointing to the unreliability of the method.”); *id.* at 6 (With respect to non-DNA hair comparison, “the acknowledged lack of any empirical evidence about false-positive rates indeed means that, as a forensic feature-comparison method, hair comparison lacks a scientific foundation. . . . PCAST concludes that there are no empirical studies that establish the scientific validity and estimate the reliability of hair comparison as a forensic feature-comparison method.”); NIST Bitmark Report, *supra* note 110, at 24 (concluding “forensic bitmark analysis lacks a sufficient scientific foundation” and stating that “the data available does not support the accurate use of bitmark analysis to exclude or not exclude individuals as the source of the bitmark”).

113. A 2018 study concluded that the method of fingerprint experts is “far from error free.” See Garrett & Mitchell, *supra* note 19, at 908 (summarizing their findings of an error rate ranging from 1–2% to 10–20% per test with respect to false positives, with an average of 7% false positives and 7% false negatives over a nearly 20-year period from 1995–2016). Several black box studies have been conducted in multiple forensic disciplines since the 2016 PCAST Report, suggesting a very low rate of error. However, those examining the studies claim that the accuracy is lower than reported, owing to a variety of causes that reflect “missingness problems”: self-selected participants, high rates of attrition, nonresponse, how inconclusive responses are used, low reproducibility (inter-examiner consistency) and repeatability (intra-examiner consistency). Khan & Carriquiry, *supra* note 102, at 3–6. The “tremendously high rates of missingness” precludes a true estimate of error rates. *Id.* at 17–18. For a critical analysis of the error rates of black box studies conducted on firearms analysis, see Heike Hofmann, Alicia Carriquiry & Susan Vanderplas, *Treatment of Inconclusives in the AFTE Range of Conclusions*, 19 L., Probability & Risk 317, 342–43 (2020), <https://doi.org/10.1093/lpr/mgab002> (addressing the effect of inconclusive findings on error rate accuracy and concluding that there is “significant work to be done” before the authors could confidently provide an error rate associated with firearms and toolmark analysis).

standards, empirical testing, or controlled methods.¹¹⁴ There were few, if any, scientists involved in the development of these specialties.¹¹⁵ Many in the forensic community were resistant to change, and historically courts were lenient in evaluating admissibility of such evidence.¹¹⁶

While courts frequently admit feature comparison evidence without limitation in many specialties,¹¹⁷ several federal courts have noted these concerns, particularly in the areas of handwriting comparison and firearm comparison.¹¹⁸ Even with respect to fingerprint analysis, Judge Frank Easterbrook, writing for the U.S. Court of Appeals for the Seventh Circuit, noted that the method's "false positive rate . . . is substantial and is likely to be higher than expected by many jurors based on longstanding claims about the infallibility of fingerprint analysis."¹¹⁹ Following the suggestions of the advisory committee's notes to the 2023 amendment to Rule 702, courts should ensure that examiner testimony is limited to what is supportable by the empirical data. Courts might choose, for example, to limit the direct testimony to the known error rate, restrict experts' certainty of conclusions, limit the language experts use, or in some cases exclude the evidence entirely.¹²⁰

Concerns about reliability as applied. According to PCAST and other commentators, one key to showing a feature or pattern method's "reliability as applied" is proficiency testing, or "ongoing empirical tests to 'evaluate the capability and performance of analysts.'"¹²¹ Such testing is particularly critical where a method rests in large part on the expert's human judgment, which is especially vulnerable to error and inter-examiner variability.¹²² Given how central an

114. See generally Jennifer L. Mnookin et al., *The Need for a Research Culture in the Forensic Sciences*, 58 U.C.L.A. L. Rev. 725 (2011) (noting the nonscientific origin of pattern disciplines and the lack of a culture of scientific inquiry in these fields).

115. Michael J. Saks & Jonathan J. Koehler, *The Coming Paradigm Shift in Forensic Identification Science*, 309 Science 892, 893 (2005), <https://doi.org/10.1126/science.1111565>.

116. 2009 NRC Report, *supra* note 37, at 108–09 (citations omitted).

117. Giannelli, *supra* note 107 (explaining that, with few exceptions, courts have continued to admit forensic science in the same way as they did pre-*Daubert*).

118. See sections titled "Handwriting Evidence" and "Firearms and Toolmark Analysis" below.

119. *United States v. Bonds*, 922 F.3d 343, 345 (7th Cir. 2019).

120. "Forensic experts should avoid assertions of absolute or one hundred percent certainty—or to a reasonable degree of scientific certainty—if the methodology is subjective and thus potentially subject to error." Fed. R. Evid. 702 advisory committee's note to 2023 amendment. The notes additionally urge that when making the decision whether to admit forensic expert testimony, "the judge should (where possible) receive an estimate of the known or potential rate of error of the methodology employed, based (where appropriate) on studies that reflect how often the method produces accurate results." *Id.*

121. 2016 PCAST Report, *supra* note 20, at 57.

122. *Id.* at 58. See also Garrett & Mitchell, *supra* note 19, at 902 (explaining why "proficiency testing is the only objective means of assessing the accuracy and reliability of experts who rely on subjective judgments to formulate their opinions (so-called 'black-box experts')").

expert's individual skill is to the success (or not) of relatively subjective methods, the demonstration of such individual skill is important, and—according to these commenters—the only way to meaningfully demonstrate such skill is by measuring each expert's successful performance on realistic proficiency tests.¹²³ One recurring challenge, however, is that existing proficiency tests tend to be relatively easy, arguably not a meaningful check on examiner competence.¹²⁴

Of course, proficiency testing alone only shows that the expert is qualified, not that the expert reliably applied the method in the case at hand. Also relevant to showing reliability as applied is documentation that the method was correctly applied in the case at hand, by the particular laboratory and examiner(s). This documentation can come in the form of internal validation, protocols, bench notes, reports, and testimony as to what the expert did and did not do in the particular case and the conditions of the sample and testing.

Moreover, to stay within the bounds of validity as applied, the examiners cannot make claims that go beyond the empirical evidence, and any limitations of the method should be acknowledged along with the demonstrated error rate of the method used.¹²⁵ The advisory committee's note to the 2023 amendment to Rule 702(d) makes clear that the expert's application of the methodology should be deemed unreliable if they overstate the strength of the inference that can be drawn from the methodology's application in a given case.¹²⁶ Expert opinion testimony regarding the weight of feature comparison evidence (i.e., evidence that a set of features corresponds between two examined items) must be limited to those inferences that can reasonably be drawn from a reliable application of the principles and methods. Thus, a statement of certainty or source attribution might not be a reliable application of a particular method, even if a statement about a class characteristic is.

To be sure, most courts are hesitant to exclude testimony where there are concerns about how well the examiner has applied the methodology to the case at hand, often finding such issues are properly matters for cross-examination.¹²⁷

123. Given the wide variability of accuracy among examiners, each expert should be performance-tested. See Bradford T. Ulery et al., *Accuracy and Reliability of Forensic Latent Fingerprint Decisions*, 108 Proc. Nat'l Acad. Scis. 7733 (2011), <https://doi.org/10.1073/pnas.1018707108> (noting the disagreement among examiners about whether fingerprints were suitable for reaching a conclusion).

124. The president of the Collaborative Testing Services (CTS) has candidly acknowledged that “[e]asy tests are favored by the community” of customers for such tests. 2016 PCAST Report, *supra* note 20, at 57 n.133 (quoting president of CTS).

125. *Id.* at 50–51. Note that calculation of the error rate might itself be controversial in a particular case, given issues related to, for example, whether to treat inconclusive determinations as false positives. See discussion at section titled “Firearms and Toolmark Analysis” below (discussing firearms validation studies).

126. See Fed. R. Evid. 702 advisory committee's note to 2023 amendment.

127. See, e.g., Itiel E. Dror, Bridget M. McCormack & Jules Epstein, *Cognitive Bias and Its Impact on Expert Witnesses and the Court*, 54 Judges J. 8, 11 (2015) (the prevailing view is that “errors

Nonetheless, there are decisions where courts have not permitted testimony where the practitioner has not reliably applied the methodology to the facts of the case.¹²⁸ In any event, the 2023 amendments to Rule 702 make clear that testimony is not admissible where the expert has not reliably applied the methodology to the case at hand. This shortcoming is not simply a matter for cross-examination but is critical to the decision about admissibility. If the judge is not satisfied by a preponderance of the evidence that the expert has reliably applied the methodology to the case at hand, the testimony should not be admitted.

Discussion of Selected Forensic Feature Comparison Evidence

Differences Between DNA and Other Methods

This reference guide addresses only non-DNA forensic feature comparison techniques, in part because DNA is ubiquitous and complex enough that it is deserving of separate treatment, and in part because of inherent differences between DNA and other existing techniques. Judges would benefit from a basic understanding of these differences.

First, forensic DNA typing was developed in the medical research context. As discussed above, most forensic disciplines were developed by law enforcement, housed in law enforcement facilities, and staffed by law enforcement personnel. In the 2009 NRC Report's view, this lack of independence from law enforcement has contributed to contextual bias, insistence upon a "zero error" rate and other unscientific claims, and a dearth of empirical studies.¹²⁹ Whatever one thinks of the NRC's conclusions in this regard, the fact is that much of the literature judges will read in forensic science litigation will tout this difference as a reason to view single-source DNA as the "gold standard" of forensic feature comparison. Second, the criteria for determining whether two DNA profiles are consistent with each other are relatively objective, precise, and repeatable compared to other feature comparison methods. While there is certainly some

in application should result in the exclusion of evidence only if they render the expert's conclusions unreliable; otherwise, the jury should be allowed to consider whether the expert properly applied the methodology in determining the weight or credibility of the expert testimony" (quoting *State v. Bernstein*, 349 P.3d 200, 201 (Ariz. 2015)).

128. See, e.g., *State v. McPhaul*, 808 S.E.2d 294, 305 (N.C. Ct. App. 2017) (holding that the trial court abused its discretion in admitting fingerprint comparison evidence where the witness failed to explain how she reached her conclusion in the case at hand).

129. See 2009 NRC Report, *supra* note 37, at 71, 99, 104. See generally Erin E. Murphy, *What "Strengthening Forensic Science" Today Means for Tomorrow: DNA Exceptionalism and the 2009 NAS Report*, 9 L., Probability & Risk 7 (2010), <https://doi.org/10.1093/lpr/mgp030> (explaining the NRC's comparison of DNA typing to other more subjective techniques).

judgment involved in determining what is and is not a “real” genetic marker, it is more straightforward to compare two graphs and note that both have a particular marker at a particular location than it is to determine that a smudged characteristic in a latent fingerprint “matches” a characteristic in a reference print. Moreover, as the *Reference Guide on Human DNA Identification Evidence*, in this manual, discusses, validation studies estimate the variability of genetic markers in the population. Third, the data underlying DNA testing allow for a quantified measure of uncertainty rather than merely a statement that two features or patterns “match” or are “consistent,” or reliance on an examiner’s judgment to decide how many similarities merit a source identification.

Ultimately, DNA will always be different from other feature comparison disciplines. DNA has high probative value in part because it is based on a biological model that explains inheritance; a statistical model that faithfully corresponds to the biology; and the population data that enables estimation of the statistical model parameters. This unique paradigm for DNA cannot be applied in the same way to other types of evidence (such as friction ridge examinations), much less evidence that is manufactured (like a toolmark). That said, other aspects of DNA typing, such as collection of population data about the variability and frequency of features and development of quantified match statistics and error rates, could be replicated in non-DNA disciplines with more research and testing.

Fingerprint Evidence

Introduction

Fingerprinting has been a forensic technique since the mid-1800s,¹³⁰ and English and American courts have accepted fingerprint identification testimony for just over a century.¹³¹ Over the years, at least before DNA, fingerprint analysis

130. Francis Galton authored the first textbook on the subject. Francis Galton, *Fingerprints* (1892) (describing the history of forensic fingerprinting, the ridges on hands and feet, and suggesting techniques for measuring and comparing ridges, including alleged racial differences). The origin of fingerprinting, like that of some other biometric techniques, is fraught in its connections to eugenics and “scientific racism,” and is far different from the discipline’s modern uses. *See generally* Simon Cole, *Suspect Identities: A History of Fingerprint and Criminal Identification* (2001). For more on the history of fingerprint identification, *see* Jennifer L. Mnookin, *Fingerprint Evidence in an Age of DNA Profiling*, 67 *Brook. L. Rev.* 13 (2001).

131. *United States v. Llera Plaza*, 188 F. Supp. 2d 549, 572 (E.D. Pa. 2002) (describing a 1906 English case in which “a New York City detective who had in 1904 been posted to Scotland Yard to learn about fingerprinting . . . used his new training to break open two celebrated cases: in each instance fingerprint identification led the suspect to confess. . . .”). The first published American appellate opinion sustaining admission of fingerprint testimony was *People v. Jennings*, 96 N.E. 1077 (Ill. 1911).

became known as the “gold standard” of forensic feature comparison expertise.¹³² In fact, proponents of new, emerging techniques in forensics would sometimes attempt to invoke onto the new techniques the prestige of fingerprint analysis.¹³³ Likewise, some early proponents of DNA typing alluded to it as “DNA fingerprinting.”¹³⁴ Fingerprint comparison is still widely used and is one of the most relied-upon forms of forensic feature comparison across the world.

This is not to say that fingerprint comparisons are free from error. In fact, the FBI’s incorrect identification of attorney Brandon Mayfield as the perpetrator of the 2004 Madrid train bombing is perhaps the most well-known error and one that spurred additional research.¹³⁵ Like other forms of forensic feature comparison, fingerprint comparison was judicially adopted without any proof of its foundational reliability. As noted below, at least one court has limited expert testimony about fingerprints and government reports have raised new questions about the limits of its accuracy.

The Method and Its Claims

Fingerprint analysis involves comparing the features of an evidence print (typically a “latent” invisible print lifted from a surface) and a reference print to determine if they might have a common source. Fingerprint comparison has historically been based on three assumptions: (1) the uniqueness of each person’s friction ridges on their fingers, (2) the permanence of those ridges throughout a person’s life, and (3) the transferability of an impression of that uniqueness to another surface. Contemporary research addresses each of these assumptions. First, scientific research has “convincingly established” that ridge patterns of humans’ fingers vary greatly among individuals and, as such, fingerprint comparison is a theoretically viable way to distinguish individuals.¹³⁶ Second, fingerprints appear to be persistent over one’s lifetime, although there can be minor changes in friction ridge detail due to aging or occupation.¹³⁷ Third, there is

132. See Sandy L. Zabell, *Fingerprint Evidence*, 13 J. L. & Pol’y 143 (2005) (noting fingerprint evidence was long considered the “gold standard” of human identification).

133. See, e.g., Kenneth Thomas, *Voiceprint—Myth or Miracle*, in *Scientific and Expert Evidence* 1015 (2d ed. 2011) (noting that advocates of sound spectrography referred to it as “voiceprint” analysis).

134. Colin Norman, *Maine Case Deals Blow to DNA Fingerprinting*, 246 Science 1556 (1989), <https://doi.org/10.1126/science.2688090>.

135. See Robert B. Stacey, *Report on the Erroneous Fingerprint Individualization in the Madrid Train Bombing Case*, Fed. Bureau of Investigation (Jan. 2005), <https://perma.cc/5WY6-AWHE>.

136. See AAAS Fingerprint Report, *supra* note 110, at 17. This variability extends to related individuals, and even identical twins develop distinguishable friction ridge skill detail. *Id.* (citations omitted). Distinguishing identical twins occurs at a slightly lower accuracy than non-twins, albeit with “relatively high accuracy.” *Id.*

137. *Id.* at 27–28 (citing numerous sources).

ample evidence that fingerprints left on a surface can be accurately transferred to another surface for purposes of comparison using various methods, although as stated, the quality (and thus analytical value) of these prints may vary dramatically from “rolled” prints (“record” prints that are typically rolled onto a fingerprint card or digitized and scanned into an electronic file).¹³⁸ For example, latent prints from a crime scene vary in pressure, and the elasticity of skin naturally distorts the impression.¹³⁹ Consequently, fingerprint impressions from the same person typically differ in some respects each time the impression is left on an object.¹⁴⁰ The latent print might sometimes be so fragmentary or smudged that analysis is impossible.

Fingerprint examiners generally follow a procedure known as analysis, comparison, evaluation, and verification (ACE-V). In the analysis stage, the examiner studies the evidence print (typically a latent print from a crime scene) to determine whether the quantity and quality of details in the print are sufficient to permit further evaluation, and determines what features are present.¹⁴¹ In the comparison stage, the examiner visually compares the evidence and known prints to determine “the details that correspond” between them, such as the “overall shape” of the prints, “lengths of the ridges, minutia location and type, thickness of the ridges and furrows, shapes of the ridges, pore position, crease patterns and shapes, scar shapes, and temporary feature shapes (e.g., a wart).”¹⁴² In the evaluation stage, the examiner “evaluates the sufficiency of the detail present to establish an identification (source determination),” determining whether, “based on his or her experience,” a “sufficient quantity and quality of friction ridge detail is in agreement between the latent print and the known

138. See David A. Stoney, *Scientific Status*, in *Modern Scientific Evidence: The Law and Science of Expert Testimony* § 32.29 (David L. Faigman et al., 2021–2022 ed.) (discussing the various methods of extraction); *United State v. Cruz-Mercedes*, 379 F. Supp. 3d 24, 44–45 (D. Mass.), *aff’d on other grounds*, 945 F.3d 569 (1st Cir. 2019) (explaining how there was only one available latent print from which to make a comparison; the remaining prints were insufficiently clear to be useful).

139. See, e.g., Stoney, *supra* note 138, § 32.32 (“Smudging, dirty surfaces, dirty fingers, and contingencies of fingermark deposition all contribute to the incomplete transfer of the finger’s detail and the introduction of specious or artifactual detail.”); *United States v. Mitchell*, 365 F.3d 215, 220–21 (3d Cir. 2004) (“Criminals generally do not leave behind full fingerprints on clean, flat surfaces. Rather, they leave fragments that are often distorted or marred by artifacts. . . . Testimony at the *Daubert* hearing suggested that the typical latent print is a fraction—perhaps 1/5th—of the size of a full fingerprint.”); *id.* at 221 n.1 (“In the jargon, artifacts are generally small amounts of dirt or grease that masquerade as parts of the ridge impressions seen in a fingerprint, while distortions are produced by smudging or too much pressure in making the print, which tends to flatten the ridges on the finger and obscure their detail.”).

140. 2009 NRC Report, *supra* note 37, at 144 (“The impression left by a given finger will differ every time, because of inevitable variations in pressure, which change the degree of contact between each part of the ridge structure and the impression medium.”).

141. *Id.* at 137–38.

142. *Id.* at 138–39. See also Stoney, *supra* note 138, § 32.32.

print” to tell whether the prints do or do not come from the same source.¹⁴³ In the *verification* stage, a second examiner repeats the analysis to see if they arrive at the same conclusion. Verification is blind if the second examiner does not know the first examiner’s conclusion.

Each of these stages in ACE-V affords examiners significant discretion. For example, examiners themselves (at least those in the United States) determine based on their judgment and experience not only how many points of comparison are sufficient before drawing a conclusion,¹⁴⁴ but whether any apparent dissimilarities (which would otherwise require reporting an “exclusion”) can be discounted as an artifact or are a result of distortion.¹⁴⁵ The examiner also has broad discretion in how to evaluate such factors as “inevitable variations” in pressure, but to date these factors have not been “characterized, quantified, or compared.”¹⁴⁶

Examiners often use the Automated Fingerprint Identification System (AFIS) of databases to rapidly search fingerprint databases in no-suspect cases. This system of databases was developed in the late 1970s and has become more important for law enforcement with the growth of databases of known prints.¹⁴⁷ AFIS is highly accurate for comparison of all ten rolled or slapped prints to another set of known prints. AFIS searches are also useful for screening large numbers of prints, which can then be evaluated by human examiners.

Scientific Assessments, Critiques, and Debates

The following discussion is focused on critiques of fingerprint analysis, given that these will frame the litigation judges will face over admission of latent print examiner testimony.

Since the 2004 Brandon Mayfield false positive FBI incident and the 2009 NRC Report, critics have focused primarily on the following concerns with

143. 2009 NRC Report, *supra* note 37, at 138.

144. Stoney, *supra* note 138, § 32.35; *United States v. Llera Plaza*, 188 F. Supp. 2d 549, 566–71 (E.D. Pa. 2002) (noting that U.S. and Scotland Yard examiners have no minimum or set number of points of comparison); 2009 NRC Report, *supra* note 37, at 141 (“The latent print community in the United States has eschewed numerical scores and corresponding thresholds” and consequently relies “on primarily subjective criteria” in making the ultimate attribution decision).

145. *Commonwealth v. Patterson*, 840 N.E.2d 12, 17 (Mass. 2005):

There is a rule of examination, the “one-discrepancy” rule, that provides that a nonidentification finding should be made if a single discrepancy exists. However, the examiner has the discretion to ignore a possible discrepancy if he concludes, based on his experience and the application of various factors, that the discrepancy might have been caused by distortions of the fingerprint at the time it was made or at the time it was collected.

146. 2009 NRC Report, *supra* note 37, at 144.

147. AAAS Fingerprint Report, *supra* note 110, at 29. AFIS includes a few million subjects for state law enforcement agencies to over a hundred million for national agencies.

respect to fingerprint analysis: its relative subjectivity, both in terms of reliance on judgment and lack of standards related to determining minimum number of points of comparison or sufficiency of detail (and thus its tendency to allow contextual bias and inter- and intra-examiner variability); the lack of variability data (about rarity of various features) to support determinations of source identification; the lack of well-designed validation studies and accurate error rate estimates; and problems with specific applications of ACE-V, such as overreliance on AFIS, failure to examine a sufficient number of details, problems with certain latent prints that go beyond the method's capabilities, lack of well-designed proficiency testing, and nonblind verification (where the verifying examiner already knows what conclusion the first examiner came to, and thus might be biased in favor of affirming the first conclusion).

Both the FBI's post-Mayfield internal investigative report and the 2009 NRC Report focused on concerns related to subjectivity, lack of variability data, and reliability as applied. The FBI noted that the Mayfield false positive result was likely the result of examiners falsely assuming that the source of the latent print was among the top potential matches suggested by AFIS, and the fact that Mayfield's print and the latent print were, in fact, remarkably similar (and thus also remarkably similar to the real source, a different suspect in Spain).¹⁴⁸ The 2009 NRC Report focused on the relative subjectivity of the method and lack of data as to how rare or common a particular type of fingerprint characteristic is.¹⁴⁹ As the 2009 NRC Report explained, "the ACE-V method does not specify particular measurements or a standard test protocol, . . . examiners must make subjective assessments throughout."¹⁵⁰ Or, as statistician Sandy Zabell put it a year after the Mayfield incident: "In contrast to the scientifically-based statistical calculations performed by a forensic scientist in analyzing DNA profile frequencies, each fingerprint examiner renders an opinion as to the similarity of friction ridge detail based on his subjective judgment."¹⁵¹ A later study by psychologist Itiel Dror and others found significant differences in latent print examiner conclusions related to the same set of samples depending on whether the examiner had been exposed to task-irrelevant information, and inconsistency within the same examiner's conclusions over time.¹⁵² The 2009 NRC Report's conclusion was that the ACE-V method, while then nearly universally accepted by courts, was too "broadly stated" to "qualify as a validated method for this type of analysis."¹⁵³

148. See generally Off. of the Inspector Gen., U.S. Dep't of Justice, A Review of the FBI's Handling of the Brandon Mayfield Case (Mar. 2006), <https://perma.cc/7UY6-575D>.

149. 2009 NRC Report, *supra* note 37, at 139–40, 144.

150. *Id.* at 139.

151. Zabell, *supra* note 132, at 158.

152. Dror et al., *supra* note 51.

153. 2009 NRC Report, *supra* note 37, at 142.

The 2016 PCAST Report's evaluation of ACE-V focused largely on (1) whether sufficient well-designed validation studies existed and what the estimated error rate was; and (2) the reliability of ACE-V as applied. The PCAST Report identified two friction ridge studies that met its criteria for a well-designed validation study, meaning one with samples representative of real case-work; a large enough sample to generate a meaningful error rate estimate; double-blind (meaning subject and grader do not realize it is a test or what the results are); and not subject to changing testing protocol midstream.¹⁵⁴ First, the FBI published a peer-reviewed study, in 2011, that was conducted in direct response to the 2009 NRC Report.¹⁵⁵ The false positive rate from the FBI study was 0.17%, with an upper bound of a 95% confidence interval of 0.33% (meaning, in simple terms, that the error rate is likely to be within a certain range with 0.33% being the highest in the range).¹⁵⁶ In terms of application, these rates correspond to a false positive occurring in an estimated 1 in every 604 cases, but potentially as often as 1 in every 306 cases.¹⁵⁷ In 2012, the FBI conducted a follow-up study directed toward assessing repeatability and reproducibility. In the follow-up study, 75 of the original 169 examiners participated, and the false positive rate was broadly consistent with the prior study.¹⁵⁸

The second and final ACE-V study identified by the PCAST Report as meeting its criteria for a well-designed validation study was done in 2014 by the Miami-Dade Police Department Forensic Services Bureau.¹⁵⁹ This study had a false positive rate of 4.2%, with an upper bound of a 95% confidence interval of 5.4%.¹⁶⁰ These rates correspond to a false positive occurring in an estimated 1 in 24 cases, but potentially as often as 1 in 18 cases.¹⁶¹ Of note, the study's authors identified 35 potential clerical errors in the documentation of the participants.¹⁶² If the 35 known errors were properly accounted for, then the actual false positive rate would have been 0.7%, with an upper bound of a 95% confidence interval of 1.4%.¹⁶³ These rates would correspond to a false positive occurring in an estimated 1 in 146

154. 2016 PCAST Report, *supra* note 20, at 94–97 (two studies); 52 (noting criteria for well-designed study).

155. *Id.* at 94.

156. *Id.*

157. *Id.*

158. *Id.* For more on the repeatability and reproducibility of examiner's decisions, see Bradford T. Ulery et al., *Repeatability and Reproducibility of Decisions by Latent Fingerprint Examiners*, 7 PLoS ONE e32800 (2012), <https://doi.org/10.1371/journal.pone.0032800>.

159. 2016 PCAST Report, *supra* note 20, at 94–95. The study can be found at <https://perma.cc/C47K-GCQA>.

160. *Id.* at 95.

161. *Id.*

162. *Id.*

163. *Id.*

cases, but potentially as often as 1 in 73 cases.¹⁶⁴ However, the report noted that to exclude errors in a post hoc manner was inappropriate for a validation study.¹⁶⁵

The 2016 PCAST Report ultimately concluded based on the two well-designed studies that “latent fingerprint analysis is a foundationally valid subjective methodology—albeit with a false positive rate that is substantial and is likely to be higher than expected by many jurors based on longstanding claims about the infallibility of fingerprint analysis.”¹⁶⁶ Note that the report’s use of the term foundationally valid appears to track the concept of “foundational reliability” in *Daubert* and its progeny. The 2016 PCAST Report did also suggest that juries in cases involving latent print expert testimony be informed that: (1) there have only been two properly conducted studies whose results are worth considering; and (2) the false positive rates from these studies “could be as high as 1 in 306 in one study and 1 in 18 in the other study.”¹⁶⁷ It also suggested that examiners acknowledge that because the participants knew they were participating in the study, the actual false positive rate could be higher than observed.¹⁶⁸ Regarding this third point, however, the report also noted, in a separate section, that “[i]t is likely that a properly designed program of systemic, blind verification would decrease the false-positive rate, because examiners in the studies tend to make *different* mistakes.”¹⁶⁹ Giving this information to factfinders, according to the report’s authors, would allow jurors “to weigh the probative value of the evidence.”¹⁷⁰

Even though it concluded latent fingerprint analysis is foundationally valid (or foundationally “reliable” in *Daubert* parlance), the 2016 PCAST Report concluded there were several issues related to its validity (or reliability, in *Daubert* parlance) *as applied*.¹⁷¹ These issues are: (1) confirmation bias—examiners altering initially marked features in a latent print based on comparison with an exemplar, (2) contextual bias—other facts of the case influencing the examiner’s judgment, and (3) lack of proficiency testing.¹⁷² Given these concerns, the report concluded the following would be necessary to establish validity of ACE-V as applied:

From a scientific standpoint, validity as applied requires that an expert:
(1) has undergone appropriate proficiency testing to ensure that he or she is

164. *Id.*

165. *Id.*

166. *Id.* at 101.

167. *Id.* at 96. The 2016 PCAST Report authors determined the number of studies worth considering based on their framework; other studies may provide helpful information. Moreover, the report authors urged that the error rates from existing validation studies be calculated based on the number of incorrect calls as a percentage of all *conclusive* examinations, and that they treat “inconclusives” as false positives. *Id.* at 93.

168. *Id.* at 109.

169. *Id.* at 96 (emphasis in original).

170. *Id.*

171. *Id.* at 102.

172. *Id.*

capable of analyzing the full range of latent fingerprints encountered in case-work and reports the results of the proficiency testing; (2) discloses whether he or she documented the features in the latent print in writing before comparing it to the known print; (3) provides a written analysis explaining the selection and comparison of the features; (4) discloses whether, when performing the examination, he or she was aware of any other facts of the case that might influence the conclusion; and (5) verifies that the latent print in the case at hand is similar in quality to the range of latent prints considered in the foundational studies.¹⁷³

Note that while proficiency testing might also be related to the qualification of a latent print examiner, the requirements above, beyond proficiency testing alone, relate to documentation of the method sufficient to support a finding by a preponderance of the evidence that the examiner reliably applied ACE-V in a particular case.

The findings of PCAST as to ACE-V's error rate and limitations were repeated in a report on latent print examination the following year by the AAAS.¹⁷⁴ One difference, however, was that the AAAS report looked at the entire body of research surrounding latent fingerprint analysis, whereas the PCAST report drew its conclusions from what it determined were the two appropriately designed black box studies.¹⁷⁵ Despite the differences in scope of review, AAAS stated that its "conclusions largely align with those of the PCAST report."¹⁷⁶

The AAAS report also mirrored the 2009 NRC Report's concern about lack of empirical data underlying source conclusion and concluded that statements of source attribution are as yet scientifically unsupportable. While the AAAS report concluded that research had "convincingly established" that ridge patterns vary greatly among humans,¹⁷⁷ it found no scientific basis for determining the rarity of any such ridge feature¹⁷⁸ or "how many features, of what types, are needed in order for an examiner to draw definitive conclusions about whether a latent print was made by a given individual."¹⁷⁹ As a result, "[e]xaminers may well be able to exclude the preponderance of the human population as possible sources of a latent print, but there is no scientific basis for estimating the number of people who could not be excluded," and there is no science to support the determination that the latent print came from a single source.¹⁸⁰ Thus, any notion that an examiner can make or has

173. *Id.*

174. AAAS Fingerprint Report, *supra* note 110.

175. *Id.*

176. *Id.* at 44.

177. *Id.* at 23.

178. *Id.*

179. *Id.* at 21.

180. *Id.*

made an identification with 100% accuracy is “overstated” and “indefensible,”¹⁸¹ as is any “claim that they can associate a latent print with a single source.”¹⁸²

The AAAS report then offered several suggestions to address any misconceptions a juror may harbor based on previous assertions of either source attribution or infallibility:

[L]atent print examiners should include specific caveats in reports that acknowledge the limitations of the discipline. They should acknowledge: (1) that the conclusions being reported are opinions rather than facts (as in all pattern-matching disciplines); (2) that it is not possible for a latent print examiner to determine that two friction ridge impressions originated from the same source to the exclusion of all others; and (3) that errors have occurred in studies of the accuracy of latent print examination.¹⁸³

The AAAS report also recommended that examiners refrain from making statements that are not supported by the science behind latent fingerprint examination.¹⁸⁴ Any words or phrases—such as “match,” “identification,” and “individualization”—that would imply a single source should be avoided.¹⁸⁵ The AAAS report recommended proposed testimony an examiner might offer that avoids any such peril:

The latent print on Exhibit ## and the record fingerprint bearing the name XXXX have a great deal of corresponding ridge detail with no differences that would indicate they were made by different fingers. There is no way to determine how many other people might have a finger with a corresponding set of ridge features, but this degree of similarity is far greater than I have ever seen in non-matched comparisons.¹⁸⁶

The Department of Justice, for its part, has changed its source conclusion language and other practices (such as nonblind verification) in response to the reports cited above. The FBI, for example, now forbids claims of “individualization.”¹⁸⁷ DOJ’s Uniform Language for Testimony and Reports (ULTR) also prohibits testifying examiners from “assert[ing] that a ‘source identification’ or a ‘source exclusion’ conclusion is based on the ‘uniqueness’ of an item of evidence,” “us[ing] the terms ‘individualize’ or ‘individualization’ when describing a source conclusion,” or “assert[ing] that two friction ridge skin impressions originated from the same

181. *Id.* at 71.

182. *Id.* at 60. U.S. Dep’t of Justice, Approved Uniform Language for Testimony and Reports for the Forensic Latent Print Discipline (2018), <https://perma.cc/E59D-B6C2>.

183. AAAS Fingerprint Report, *supra* note 110, at 73.

184. *Id.* at 11.

185. *Id.*

186. *Id.*

187. See Fed. Bureau of Investigation, FBI Approved Standards for Scientific Testimony and Report Language for Forensic Document Comparisons 4 (2022), <https://perma.cc/VZ5C-5ZBT>.

source to the exclusion of all other sources.”¹⁸⁸ These prohibitions are in addition to the proscription of asserting latent fingerprint analysis is “infallible or has a zero error rate.”¹⁸⁹

Notwithstanding the recent purging of terms like individualization and zero error rate, nearly all laboratories, including DOJ’s, still allow examiners to testify to language suggesting categorical source attribution or exclusion. DOJ’s ULTR now directs their examiners to arrive at one of three conclusions: a source identification, a source exclusion, or an inconclusive determination.¹⁹⁰ A source identification is an examiner’s opinion that the two prints came from the same person and is based on an examiner’s opinion that there is “extremely strong support” indicating the two prints came from the same source and “extremely weak support” indicating the two prints came from different sources.¹⁹¹ A source exclusion is an examiner’s opinion that the two prints did not come from the same source and is based on the examiner’s opinion that there is “extremely strong support” indicating the two prints came from different sources and “extremely weak or no support” that the two prints came from the same source.¹⁹² An examiner will make an inconclusive determination when “there is insufficient quantity and/or clarity” between the two impressions for the examiner to arrive at a source identification or source exclusion determination.¹⁹³ While these terms do not claim infallibility, they do suggest the ability to reliably conclude that prints have a single common source. Concerns still linger, of course, about whether jurors understand the significance of such limitations.¹⁹⁴

A final recurring concern relates to use of AFIS databases. The Mayfield case was one high-profile example of a false hit from a database search. As databases grow, they may contain more fingerprints with many common features and few discernible dissimilar features, known as close non-“matches” (CNMs). In a study evaluating 125 fingerprint agencies completing a mandatory proficiency test with two pairs of CNMs, the false positive rate was 15.9% and 28.1% respectively, raising serious concerns.¹⁹⁵ While there are continued efforts to

188. U.S. Dep’t of Justice (DOJ), Uniform Language for Testimony and Reports for the Forensic Latent Print Discipline 3 (2020), <https://perma.cc/J7GQ-US5Y>.

189. *Id.*

190. *Id.* at 2.

191. *Id.* In other words, “[a]n identification is the statement of an examiner’s opinion (an inductive inference) that the probability that the two impressions were made by different sources is so small that it is negligible” (internal citation omitted).

192. *Id.*

193. *Id.* at 3.

194. In a study in the firearms context evaluating juror comprehension about differences in expert’s conclusion language, the authors found that more modest phrasing about conclusions did not affect jurors’ conclusions. See Brandon L. Garrett et al., *Mock Jurors’ Evaluation of Firearm Expert Testimony*, 44 Law & Hum. Behav. 412 (2020), <https://doi.org/10.1037/lhb0000423>.

195. See Jonathan J. Koehler & Shiquan Liu, *Fingerprint Error Rate on Close Non-matches*, 66 J. Forensic Scis. 129, 131–32 (2020), <https://doi.org/10.1111/1556-4029.14580>. The authors

improve AFIS,¹⁹⁶ the systems are not currently designed to prove that a particular pair of prints has a common source.¹⁹⁷

Case Law Development

Since the Illinois Supreme Court's approval of fingerprint analysis in *People v. Jennings* (1911), fingerprint testimony has been routinely admitted.¹⁹⁸ Indeed, some courts stated that fingerprint evidence was the strongest proof of a person's identity.¹⁹⁹ With the exception of one federal district court decision that was later withdrawn by the court itself upon reconsideration, and another decision in which a court remanded to determine whether ACE-V might be unreliable as applied to simultaneous impressions (rather than single latent prints),²⁰⁰ nearly all courts before 2017 continued to reject challenges to fingerprint testimony.²⁰¹

note that there were limitations to the study; the participants knew they were being tested and the samples were "convenience samples," not random and representative selections from database searches. For more on AFIS searches, see Itiel E. Dror & Jennifer L. Mnookin, *The Use of Technology in Human Expert Domains: Challenges and Risks Arising from the Use of Automated Fingerprint Identification Systems in Forensic Science*, 9 Law, Probability & Risk 47 (2010), <https://doi.org/10.1093/lpr/mgp031>.

196. AAAS Fingerprint Report, *supra* note 110, at 33, quoting the 2016 PCAST Report, *supra* note 20.

197. AAAS Fingerprint Report, *supra* note 110, at 33.

198. *People v. Jennings*, 96 N.E. 1077 (Ill. 1911). As Professor Mnookin has noted, "fingerprints were accepted" after *Jenkins* "as an evidentiary tool without a great deal of scrutiny or skepticism." Mnookin, *supra* note 130, at 17.

199. *People v. Adamson*, 165 P.2d 3, 12 (Cal. 1946), *aff'd*, 332 U.S. 46 (1947).

200. *See, e.g., Commonwealth v. Patterson*, 840 N.E.2d 12, 15, 16–17 (Mass. 2005) ("These latent print impressions are almost always partial and may be distorted due to less than full, static contact with the object and to debris covering or altering the latent impression"; "In the evaluation stage, . . . the examiner relies on his subjective judgment to determine whether the quality and quantity of those similarities are sufficient to make an identification, an exclusion, or neither."); and *United States v. Llera Plaza*, 179 F. Supp. 2d 492, 516, *vacated, mot. granted on recons.*, 188 F. Supp. 2d 549 (E.D. Pa. 2002). The ruling excluded expert testimony that two sets of prints "matched," to the exclusion of all other persons. On a motion for reconsideration, the court reversed itself. A spate of legal articles followed. *See, e.g.,* Simon A. Cole, *Grandfathering Evidence: Fingerprint Admissibility Rulings from Jennings to Llera Plaza and Back Again*, 41 Am. Crim. L. Rev. 1189 (2004); Robert Epstein, *Fingerprints Meet Daubert: The Myth of Fingerprint "Science" Is Revealed*, 75 S. Calif. L. Rev. 605 (2002); Kristin Romandetti, *Recognizing and Responding to a Problem with the Admissibility of Fingerprint Evidence Under Daubert*, 45 Jurimetrics 41 (2004).

201. *See United States v. Baines*, 573 F.3d 979, 990 (10th Cir. 2009) ("Fingerprint identification has been used extensively by law enforcement agencies all over the world for almost a century."); *United States v. Abreu*, 406 F.3d 1304, 1307 (11th Cir. 2005) ("We agree with the decisions of our sister circuits and hold that the fingerprint evidence admitted in this case satisfied *Daubert*."); *United States v. Janis*, 387 F.3d 682, 690 (8th Cir. 2004) (finding fingerprint evidence to be reliable); *United States v. Mitchell*, 365 F.3d 215, 234–52 (3d Cir. 2004); *United States v. Crisp*, 324 F.3d 261, 268–71 (4th Cir. 2003); *United States v. Collins*, 340 F.3d 672, 682 (8th Cir.

Since the AAAS Fingerprint Report in 2017, at least one court has limited latent print testimony in new ways on “reliability as applied” grounds. For example, the North Carolina Supreme Court in 2018 held that the trial court abused its discretion in allowing a latent print examiner to testify to a “match” between the defendant’s known prints and a latent print on a truck, where the testimony merely established that the examiner used ACE-V, and determined the “match” based on her “training and experience” and “looking at the individual minutia” in each print.²⁰² The court found a similar error in another case two years later.²⁰³

Still, since the 2016 PCAST Report, most courts have deemed concerns with latent print analysis as going to weight, not admissibility,²⁰⁴ citing the long-standing nature of the discipline and its traditional acceptance by courts under *Daubert*.²⁰⁵ Other courts upholding the admissibility of fingerprint analysis have noted that while reports have exposed flaws in ACE-V, these reports have also opened up new opportunities for defense. One court noted that “[c]ross-examination on issues like error rates is possible now in a way in which it was not 30 years ago.”²⁰⁶ Still other courts have reasoned that even with flaws, fingerprint analysis is likely better than older evidence like eyewitness identifications and “grainy” photographs, suggesting that the middle ground is to admit such evidence and “subject [it] to cross-examination about a method’s reliability and whether the witness took appropriate steps to reduce errors.”²⁰⁷

2003) (“Fingerprint evidence and analysis is generally accepted.”); *United States v. Hernandez*, 299 F.3d 984, 991 (8th Cir. 2002); *United States v. Sullivan*, 246 F. Supp. 2d 700, 704 (E.D. Ky. 2003); *United States v. Martinez-Cintrón*, 136 F. Supp. 2d 17, 20 (D.P.R. 2001).

202. *State v. McPhaul*, 808 S.E.2d 294, 304–05 (N.C. Ct. App. 2017), *discretionary review improvidently allowed*, 818 S.E.2d 102 (N.C. 2018) (abuse of discretion to admit where the expert failed to demonstrate that she applied the principles and methods reliably to the facts of the case).

203. *Cf. State v. Koiyan*, 841 S.E.2d 351 (N.C. Ct. App. 2020) (same error as in *McPhaul*, though declining to reverse under plain error review).

204. *See United States v. Reyes-Ballista*, No. CV 18–634–2 (ADC), 2020 WL 6822372, at *4 (D.P.R. Nov. 20, 2020) (“considering that defendant will have ample opportunity to conduct vigorous cross-examination of the government’s expert witnesses and present contrary evidence, defendant is not without means of attacking the evidence he now claims to be based on methods that run afoul the profession’s parameters and accepted methods”).

205. *United States v. Pitts*, No. 16–CR–550 (DLI), 2018 WL 1116550, at *4–6 (E.D.N.Y. Feb. 26, 2018), citing *United States v. Avitia-Guillen*, 680 F.3d 1253, 1260 (10th Cir. 2012) (“Fingerprint comparison is a well-established method of identifying persons, and one we have upheld against a *Daubert* challenge.”). *See also Reyes-Ballista*, 2020 WL 6822372, at *3 (noting that “several . . . Circuits have explicitly determined that, ‘in the context of fingerprint evidence, a *Daubert* hearing is not always required’”) (internal citation omitted); *United States v. Stevens*, 219 F. App’x 108, 109 (2d Cir. 2007) (summary order declining to hold a *Daubert* hearing); *United States v. Bonds*, No. 15 CR 573–2, 2017 WL 4511061 (N.D. Ill. Oct. 10, 2017), *aff’d*, 922 F.3d 343 (7th Cir. 2019).

206. *United States v. Fell*, No. 5:01–CRCR–12–01, 2016 WL 11550830, at *1 (D. Vt. Dec. 29, 2016).

207. *Bonds*, 922 F.3d at 346 (Easterbrook, J.).

Given the 2023 amendment to Rule 702, courts may now need to engage in more robust gatekeeping to evaluate both the foundational reliability and the reliability as applied of latent print expert testimony. As the advisory committee's notes clarify, "many courts have held that the critical questions of the sufficiency of an expert's basis, and the application of the expert's methodology, are questions of weight and not admissibility. These rulings are an incorrect application of Rules 702 and 104(a)."²⁰⁸ The notes explain that "[j]udicial gatekeeping is essential." As explained earlier, several gatekeeping options are available to the courts, along with the option of providing jury instructions to aid jurors in evaluating the evidence. With the enactment of the 2023 amendment to Rule 702, the judge must be satisfied that the proponent of the evidence has met each of the rule's requirements by a preponderance. If the court is not satisfied, it should exclude those opinions. Additionally, the judge may appropriately limit any opinion to exclude overstatement of a conclusion.

Handwriting Evidence

Introduction

Individuals who compare handwriting, ink formulations, and other tasks related to evaluation of questioned documents are known as questioned document examiners or forensic document examiners (FDEs). FDEs are called on to perform a variety of tasks such as determining potential authorship of a writing; determining the sequence of strokes on a page; and determining whether a particular ink formulation existed on the purported date of a writing.²⁰⁹ Courts were originally skeptical of handwriting comparison and dismissed its value, but by 1900, many courts had begun to admit such expert evidence.²¹⁰ In the 1936 trial about the Lindbergh baby kidnapping, expert testimony about the handwriting of the random notes figured prominently.²¹¹ Following the Lindbergh

208. Fed. R. Evid. 702 advisory committee's note to 2023 amendment.

209. Questioned document examinations cover a wide range of analyses: handwriting, hand printing, typewriting, mechanical impressions, altered documents, obliterated writing, indented writing, and charred documents. See Paul C. Giannelli et al., *Scientific Evidence* § 14 (6th ed. 2020).

210. See, e.g., D. Michael Risinger, Mark P. Denbeaux, & Michael J. Saks, *Exorcism of Ignorance as a Proxy for Rational Knowledge: The Lessons of Handwriting Identification "Expertise,"* 137 U. Pa. L. Rev. 731, 762 (1989). For a deep and nuanced analysis of the history of handwriting identification, see Jennifer L. Mnookin, *Scripting Expertise: The History of Handwriting Identification Evidence and the Judicial Construction of Reliability*, 87 Va. L. Rev. 1723 (2001).

211. Risinger et al., *Exorcism of Ignorance*, *supra* note 210, at 771.

case, handwriting evidence became widely used and judicially accepted, but with little critical scrutiny by the courts into its accuracy.²¹²

Reliance on handwriting comparison has diminished in recent years and forensic document examination units are closing as demand shifts to other forensic disciplines, such as DNA analysis.²¹³ Additionally, there is a continual decrease in handwritten communication, as individuals use more electronic communication and digital signatures. These developments suggest that this field may soon join other receding identification specialties. Nevertheless, it continues to be used in court, and in 2020 a NIST working group issued a report, discussed below, for improving the practice.

The Method and Its Claims

Some FDE tasks, such as ink identification, are based on sophisticated techniques such as chromatography and mass spectrometry that are widely considered to be reliable.²¹⁴ Other tasks, such as determining authorship of a document, present more substantial potential legal issues. This section will focus primarily on authorship.

Like other feature comparison experts, an FDE determines authorship by analyzing a questioned document and known sample (either previous writing or a requested handwriting exemplar), observing various details in each, comparing them in an attempt to discern consistent or inconsistent patterns, and then evaluating those similarities and differences to arrive at a conclusion about whether they might have a common author. There must be a sufficient quantity and quality of material to permit the examiner to compare the samples.²¹⁵ The 2020 NIST Report suggests that it is best if the exemplars are written during the same period as the questioned document, to remove potential variables. In performing their comparison, examiners consider letter formation, size, and inter-word and intra-word spacing.²¹⁶

The so-called conventional or classical approach is a two-stage process. Examiners consider (1) class and (2) so-called “individual” characteristics. Of characteristics labeled “class” characteristics, two types are weighed: “system” and “group.”²¹⁷ People exhibiting system characteristics would include, for

212. *Id.* (noting that the Lindbergh case seemed to have “stamped out virtually all manifestations of judicial skepticism”).

213. 2020 NIST Report, *supra* note 110, at xi.

214. Giannelli et al., *supra* note 209, § 14.04 [4].

215. 2020 NIST Report, *supra* note 110, at 45 (noting the assumption of individuality rests on a large enough quality sample).

216. *Id.* at 14.

217. See James A. Kelly, *Questioned Document Examination*, in *Scientific and Expert Evidence* 695, 698 (2d ed. 2011).

example, those who learned the Palmer method of cursive writing, taught in many schools. People taught the same system might reasonably be expected to manifest some of the characteristics of that writing style. However, because schools now devote much less time to teaching handwriting as a skill, the more contemporary view is that the “class” characteristics are not as readily identifiable as they once were.²¹⁸ Other “group” characteristics might include those with a medical condition affecting their handwriting.²¹⁹

The conventional belief in individual characteristics unique to a particular person also persists among many examiners. That belief rests on two premises: (1) that no two writers share the same combination of writing characteristics; and (2) adults have consistent writing habits.²²⁰ These beliefs remain unproven,²²¹ but still assumed by many practitioners and still cited by some courts.²²² The traditional process for handwriting identification involves measuring selected features in handwriting, determining whether and how they differ across specimens, and interpreting the significance of similarities and differences.²²³ Although this process can include measuring features, more often it involves an examination of relative measurements (an estimation of features in proportion to each other), including size, spacing, and slant of features. Although there are differences among the various examiners, most follow these general procedures:

1. analyzing the features of the questioned writing and known standards both macroscopically and microscopically;
2. noting conspicuous features such as size, slant, and letter construction, as well as more subtle characteristics such as pen direction, the nature of connections between letters, and spacing between letters, words, and lines;
3. comparing the observed features to determine similarities and dissimilarities;
4. taking into account the degree of similarity or otherwise and the nature of the writing (quality, amount, and complexity), evaluating the evidence,

218. 2020 NIST Report, *supra* note 110, at 7–8.

219. See, e.g., Larry F. Stewart, *The Process of Forensic Handwriting Examinations*, 4 Forensic Res. & Criminology Int'l J. 139 (2017), <https://doi.org/10.15406/frcij.2017.04.00126>.

220. 2020 NIST Report, *supra* note 110, at 8 (citing Howard Sieden & Frank Norwitch, *Questioned Documents*, in *Forensic Science: An Introduction to Scientific and Investigative Techniques* (S.H. James et al. eds., 4th ed. 2014)).

221. For more, see Andrew Sulner, *Critical Issues Affecting the Reliability and Admissibility of Handwriting Opinion Evidence—How They Have Been Addressed (or Not) Since the 2009 NAS Report, and How They Should Be Addressed Going Forward: A Document Examiner Tells All*, 48 Seton Hall L. Rev. 631, 639 (2018) (“absolutist statements concerning uniqueness and intra-writer variability are as yet unproven, and likely unprovable”).

222. See, e.g., *United States v. Mallory*, 902 F.3d 584 (6th Cir. 2018); *United States v. Foust*, 989 F.3d 842, 846 (10th Cir.), *cert. denied*, 142 S. Ct. 294 (2021).

223. 2020 NIST Report, *supra* note 110, at 8.

and arriving at an opinion regarding the writership of the questioned writing.²²⁴

Examiners believe that the various features they examine, such as letter formation and size, and the inter- and intra-word spacing, are more variable in the writing of different individuals than in the writing of the same individual, but today there is widespread acknowledgment that the statistical properties of such variabilities have not been rigorously studied.²²⁵ There is no universally accepted number of points of similarity required to conclude the writings came from the same source. Additionally, although there is an expected range of variability for a single writer, there are no specific standards or procedures for determining that range.²²⁶

After evaluation, the examiner “expresses an opinion indicating [their] subjective confidence in the process outcome.”²²⁷ There are five opinions possible, according to the 2020 NIST Report: (1) Identification; (2) Probably did write; (3) Inconclusive; (4) Probably did not write; and (5) Elimination.²²⁸ In contrast, the Scientific Working Group for Forensic Document Examination (SWGDOC) gives examiners nine options that may be expressed, along with associated descriptions.²²⁹ For its part, the FBI laboratory uses five categories that “collapse” the nine possible opinions on the SWGDOC scale.²³⁰ As of 2021, DOJ requires examiners to offer any of the following five conclusions: (1) Source Identification (i.e., identified); (2) Support for a Common Source; (3) Inconclusive; (4) Support for Different Sources; and (5) Source Exclusion (i.e., excluded).²³¹ DOJ has also limited the certainty that experts may use, consistent with other forensic feature comparison methods.²³² In short, just as there is no consensus about the number of specific points of similarity to determine authorship, there is no nationwide consensus about the range of possible opinions.

224. *Id.* at 9.

225. *Id.* at 14.

226. *Id.*

227. *Id.* at 24.

228. *Id.*

229. See, e.g., 2009 NRC Report, *supra* note 37, at 166; 2020 NIST Report, *supra* note 110, at 25–26, listing: 1. Identification; 2. Strong probability; 3. Probable; 4. Indications; 5. No conclusion; 6. Indications did not; 7. Probably did not; 8. Strong probability did not; and 9. Elimination.

230. 2020 NIST Report, *supra* note 110, at 25, citing SWGDOC, Version 2013–2.

231. U.S. Dep’t of Justice, Uniform Language for Testimony and Reports for Testimony and Reports for Forensic Document Examination 2 (2021), <https://perma.cc/KMC4-6KG3>.

232. *Id.* at 4 (“Therefore, an examiner shall not: assert that a ‘source identification’ or a ‘source exclusion’ conclusion is based on the ‘uniqueness’ of an item of evidence; use the terms ‘individualize’ or ‘individualization’ when describing a source conclusion; assert that two or more bodies of writing were prepared by the same writer to the exclusion of all other writers.”).

Scientific Assessments, Critiques, and Debates

The following discussion is focused on critiques of handwriting analysis, given that these will frame the litigation judges will face over admission of FDE testimony.

As is the case with other feature comparison methods, critiques of the foundational reliability of handwriting analysis have focused on its relative subjectivity, lack of empirical data underlying assumptions about rarity of characteristics, potential for contextual bias to affect examinations, and lack of accurate error rate estimates. The 2009 NRC Report concluded that the technique was useful but not yet scientifically valid for determining authorship:

The scientific basis for handwriting comparisons needs to be strengthened. Recent studies have increased our understanding of the individuality and consistency of handwriting . . . and suggest that there may be a scientific basis for handwriting comparison, at least in the absence of intentional obfuscation or forgery. Although there has been only limited research to quantify the reliability and replicability of the practices used by trained document examiners, the committee agrees that there may be some value in handwriting analysis.²³³

A long-time forensic document examiner himself agreed with this assessment, writing in a 2018 article that the method could be very reliable under the right conditions and with more research but that “forensic handwriting analysis still lacks robust, ground truth studies that provide empirical support for the reliability of many of the tasks routinely performed.”²³⁴

The latest governmental evaluation of handwriting analysis, the 2020 NIST Report, focused its foundational reliability concerns on the relative subjectivity of the method, the resulting potential for contextual bias, and the need for empirical testing to determine the extent of both intra- and inter-examiner variability (that is, repeatability and reproducibility, two aspects of what is often called reliability, as compared to accuracy).²³⁵ To provide trustworthy estimates of repeatability and reproducibility, studies should “compare the performance within and between FDEs in their judgments on the same samples of handwriting against ground truth.”²³⁶ The report urges that multiple independent laboratories work together on these problems by using the same methods and materials.²³⁷ The NIST report ultimately urges that both “black box studies” (to calculate an accurate error rate estimate) and “white box studies” (to understand the steps examiners are taking, with an eye toward developing standards that then must be applied when shown to lead to accurate results) are needed, and

233. 2009 NRC Report, *supra* note 37, at 166–67.

234. Sulner, *supra* note 221, at 714.

235. 2020 NIST Report, *supra* note 110, at 52.

236. *Id.* at 53.

237. *Id.*

at least one has been performed since 2020 (when the NIST Human Factors report was released).²³⁸

The 2020 NIST Report also expressed concern over the method's reliability as applied, and in particular, the use by examiners of language that suggests source attribution or that otherwise goes beyond the scientifically supportable bounds of the method or overstates the evidence. The report cautioned, for example, that the expressions "uniqueness" and "individualization" do not have a settled meaning in forensic science, and their casual use can lead to misunderstandings in court and an "exaggeration of the strength of such evidence."²³⁹ The report also concluded that "empirical research and statistical reasoning do not support source attribution to the exclusion of all others." Instead it forcefully recommended that FDEs "must not report or testify, directly or by implication, that questioned handwriting has been written by an individual (to the exclusion of all others)."²⁴⁰ It likewise recommended that language about uniqueness and individualization should give way to a critical evaluation of the "rarity of the features," on a continuum of rare to common. "A more contemporary view . . . individuality is defined with respect to the probability of observing writing profiles of two individuals that are indistinguishable using the specified comparison method."²⁴¹ Additionally, there is movement away from the conventional approach to handwriting analysis and toward a more empirical, neurobiological approach that would permit hypothesis generation and testing of relevant principles.²⁴²

Finally, with respect to reliability as applied, the 2020 NIST Report dedicated a full section to the several types of cognitive bias that could arise in FDE because of its layers of relatively high subjectivity.²⁴³ The report cautioned that such bias is a "legitimate cause for concern" that must be addressed by forensic

238. See, e.g., R. Austin Hicklin et al., *Accuracy and Reliability of Forensic Handwriting Comparisons*, 119 Proc. Nat'l Acad. Scis. e2119944119 (2022), <https://doi.org/10.1073/pnas.2119944119>.

239. 2020 NIST Report, *supra* note 110, at 47.

240. *Id.* DOJ's ULTR for Forensic Document Examination continues to permit an examiner to offer a conclusion on "source identification" but must not assert that "two or more bodies of writing were prepared by the same writer to the exclusion of all other writers." U.S. Dep't of Justice, Uniform Language for Testimony and Reports for Testimony and Reports for Forensic Document Examination 3 (2021), <https://perma.cc/KMC4-6KG3>.

241. 2020 NIST Report, *supra* note 110, at 46 (citing Sargur Srihari et al., *Individuality of Handwriting*, 47 J. Forensic Scis. 856 (2002)).

242. See, e.g., Michael P. Caligiuri & Linton A. Mohammed, *The Neuroscience of Handwriting* 35–57 (2012).

243. 2020 NIST Report, *supra* note 110, at 30. The NIST report dedicates an entire section (2.1) to the multiple types of contextual bias in FDE. *Id.* at 30–44. See also D. Michael Risinger et al., *The Daubert/Kumho Implications of Observer Effects in Forensic Science: Hidden Problems of Expectation and Suggestion*, 90 Calif. L. Rev. 1 (2002), <https://doi.org/10.2307/3481305>; Adele Quigley-McBride et al., *A Practical Tool for Information Management in Forensic Decisions: Using Linear Sequential Unmasking-Expanded (LSU-E) in Casework*, 4 Forensic Sci. Int'l: Synergy 100216 (2022), <https://doi.org/10.1016/j.fsisy.2022.100216>.

laboratories through standards, protocols like the “contextual information management” designed to control information flow to and among examiners, and documentation of steps taken to avoid such bias.²⁴⁴ The report made several recommendations along these lines and urged research to study the problem.²⁴⁵ There are different suggested approaches to combat these problems—notably Dr. Itiel Dror’s hierarchy of expert performance method (HEP).²⁴⁶

Another legal question that has arisen is whether FDE expert testimony is “helpful to the jury” for Rule 702 purposes. Forensic handwriting analysis is a unique discipline because handwriting is a fundamental aspect of everyday life, and thus, ordinary people are capable of comparing some aspects of handwriting samples.²⁴⁷ Laypersons are usually capable of identifying gross characteristics, but forensic handwriting analysis focuses on nuanced details that laypersons are often unable to correctly identify.²⁴⁸ A 2022 FBI-funded study found that adequately trained forensic handwriting experts are more accurate than those with minimal training and that examiners with less than two years of formal training had higher error rates and were more likely to make definitive, unqualified conclusions compared to those with at least two years of formal training.²⁴⁹ A 2018 study reached a similar conclusion when comparing novice and expert examiners, but found that the overall error rate among expert examiners was still high enough to question the trustworthiness of handwriting comparison evidence in the courts.²⁵⁰

244. 2020 NIST Report, *supra* note 110, at 34.

245. *Id.* at 35–36. For a detailed discussion of how the contextual information management (CIM) should work in FDE, see *id.* at 36–44.

246. See, e.g., Itiel Dror, *A Hierarchy of Expert Performance*, 5 J. Applied Res. Memory & Cognition 121 (2016), <https://doi.org/10.1016/j.jarmac.2016.03.001> (creating a hierarchy of expert performance to identify weaknesses in an individual’s performance and compare experts across domains).

247. 2020 NIST Report, *supra* note 110, at viii.

248. See, e.g., *State v. Cooke*, 914 A.2d 1078, 1101 (Del. Super. Ct. 2007) (“Handwriting comparison is beyond a lay person’s general knowledge.”); Diane Harrison, Ted M. Burkes & Danielle P. Sieger, *Handwriting Examination: Meeting the Challenges of Science and the Law*, Forensic Sci. Comm’n’s (Oct. 2009), <https://perma.cc/J2Z2-W7D5>.

249. R. Austin Hicklin et al., *Accuracy and Reliability of Forensic Handwriting Comparisons*, 119 Proc. Nat’l Acad. Scis. e2119944119 (2022), <https://doi.org/10.1073/pnas.2119944119>.

250. Kristy Martire et al., *What Do the Experts Know? Calibration, Precision, and the Wisdom of Crowds Among Forensic Handwriting Experts*, 25 Psych. Bull. Rev. 2346, 2353 (2018), <https://doi.org/10.3758/s13423-018-1448-3>.

On the one hand, there is some evidence that handwriting experts will be able to estimate the frequency of occurrence for handwriting features better than novices. However, even the single best performing participant produced an average deviation of 18.5% from the true value. On the other hand, this number is considerably lower than would be expected by chance (25%).

Some examiners incorporate handwriting pattern recognition technology to assist in their analysis.²⁵¹ This technology can outperform laypersons²⁵² and can be as accurate as an expert examiner, depending on the complexity of the task.²⁵³ These systems will only become more advanced as machine learning algorithms enhance their accuracy and reliability.²⁵⁴ Another area of research has been computer-driven forensics linguistics to identify authors.²⁵⁵ For instance, a Swiss company, OrphAnalytics SA, recently released a software package for conducting stylometric analysis to support the attribution of text to a particular author. While promising, these techniques have yet to be used forensically, and as with other forms of artificial intelligence, they too will have shortcomings. The 2020 NIST Report calls for FDEs to integrate these tools into their casework and “collaborate with the computer science and engineering communities to develop and validate applicable, user-friendly, automated systems.”²⁵⁶ Training programs requiring proficiency in the use of these systems, combined with interdisciplinary research and development, should provide courts with additional metrics to evaluate accuracy and reliability of handwriting comparisons in the future.

Case Law Development

Although nineteenth-century courts were skeptical of handwriting expertise,²⁵⁷ the twentieth century saw a marked shift, as testimony in leading cases like the Lindbergh kidnapping helped the discipline gain judicial acceptance.²⁵⁸ There was little dispute that handwriting testimony was admissible at the time the

251. 2020 NIST Report, *supra* note 110, at 68–71.

252. Harrison et al., *supra* note 248 (citing Sargur Srihari et al., *On the Discriminability of the Handwriting of Twins*, 53 J. Forensic Scis. 430 (2008), <https://doi.org/10.1111/j.1556-4029.2008.00682.x>).

253. 2020 NIST Report, *supra* note 110, at 70–71 (citing Muhammad Imran Malik et al., *Man vs. Machine: A Comparative Analysis for Signature Verification*, 24 J. Forensic Document Examination 21 (2004)).

254. 2020 NIST Report, *supra* note 110, at 71.

255. See, e.g., Janet Ainsworth & Patrick Juola, *Who Wrote This?: Modern Forensic Authorship Analysis as a Model for Valid Forensic Science*, 96 Wash. U. L. Rev. 1159, 1169–72 (2019) (language has many “objectively identifiable features”); Patrick Juola, *Verifying Authorship for Forensic Purposes: A Computational Protocol and its Validation*, 325 Forensic Sci. Int’l 110824 (2021), <https://doi.org/10.1016/j.forsciint.2021.110824> (claiming a measured accuracy of 77% across more than 32,000 different document pairs); Cami Fuglsby et al., *Elucidating the Relationships Between Two Automated Handwriting Feature Quantification Systems for Multiple Pairwise Comparisons*, 67 J. Forensic Scis. 642 (2022), <https://doi.org/10.1111/1556-4029.14914>.

256. 2020 NIST Report, *supra* note 110, at 72.

257. See *Strother v. Lucas*, 31 U.S. 763, 767 (1832); *Phoenix Fire Ins. Co. v. Philip*, 13 Wend. 81, 82–84 (N.Y. Sup. Ct. 1834).

258. See Risinger et al., *Exorcism of Ignorance*, *supra* note 210, at 771.

Federal Rules of Evidence were enacted in 1975. Rule 901(b)(3) recognized that a document could be authenticated by an expert, and the drafters explicitly mentioned handwriting comparison through the “testimony of expert witnesses.”²⁵⁹

The first significant admissibility challenge under *Daubert* was in *United States v. Starzecpyzel*.²⁶⁰ In that case, the district court concluded that “forensic document examination, despite the existence of a certification program, professional journals and other trappings of science, cannot, after *Daubert*, be regarded as ‘scientific . . . knowledge.’”²⁶¹ Nevertheless, the court did not exclude handwriting comparison testimony. Instead, the court admitted the individuation testimony as nonscientific “technical” evidence.²⁶²

Starzecpyzel prompted more litigation that questioned the lack of empirical validation in the field.²⁶³ For many years, there was a three-way split of authority. The majority of courts permitted examiners to express individuation opinions.²⁶⁴ As one court noted, “all six circuits that have addressed the admissibility of handwriting expert [testimony] . . . [have] determined that it can satisfy the reliability threshold” for nonscientific expertise.²⁶⁵ In 2021, the Tenth Circuit, in *United States v. Foust*, held that handwriting comparison was properly admitted, relying largely on the “widespread acceptance of handwriting comparison through the years” and citing the importance of the general acceptance factor to its decision.²⁶⁶ The *Foust* opinion explained that the proposed testimony fell short in the “standards, peer review, and error rate” factors, but upheld the trial

259. Fed. R. Evid. 901(b)(3) advisory committee’s note.

260. 880 F. Supp. 1027 (S.D.N.Y. 1995).

261. *Id.* at 1038.

262. *Id.* at 1047.

263. See, e.g., *United States v. Hidalgo*, 229 F. Supp. 2d 961, 967 (D. Ariz. 2002):

Because the principle of uniqueness is without empirical support, we conclude that a document examiner will not be permitted to testify that the maker of a known document is the maker of the questioned document. Nor will a document examiner be able to testify as to identity in terms of probabilities.

264. See, e.g., *United States v. Prime*, 363 F.3d 1028, 1033 (9th Cir. 2004); *United States v. Crisp*, 324 F.3d 261, 265–71 (4th Cir. 2003); *United States v. Jolivet*, 224 F.3d 902, 906 (8th Cir. 2000) (affirming the introduction of expert testimony that it was likely that the accused wrote the questioned documents); *United States v. Velasquez*, 64 F.3d 844, 848–52 (3d Cir. 1995); *United States v. Ruth*, 42 M.J. 730, 732 (A. Ct. Crim. App. 1995), *aff’d on other grounds*, 46 M.J. 1 (C.A.A.F. 1997); *United States v. Morris*, No. 06–87–DCR, 2006 WL 2054585, *2 (E.D. Ky. July 20, 2006); *Orix Fin. Servs. v. Thunder Ridge Energy, Inc.*, No. 01 Civ. 4788, 2006 WL 587483 (S.D.N.Y. Mar. 8, 2006).

265. *Prime*, 363 F.3d at 1034.

266. *United States v. Foust*, 989 F.3d 842, 846 (10th Cir.), *cert denied*, 142 S. Ct. 294 (2021). For more about courts’ reliance on a “long history of use” as justification to admit forensic evidence, see Jane Campbell Moriarty, *Deceptively Simple: Framing, Intuition, and Judicial Gatekeeping of Forensic Feature-Comparison Methods Evidence*, 86 Fordham L. Rev. 1687 (2018).

judge's decision to admit the evidence. Explaining that the *Daubert* factors were meant to be helpful and not definitive, the Tenth Circuit held that the trial court did not abuse its discretion in admitting the evidence. Echoing *Starzecpyzel*, the court noted “handwriting comparison is not a traditional science.”²⁶⁷ So, too, the Sixth Circuit in *United States v. Mallory* held that handwriting comparison was properly admitted, as it was based on “knowledge and experience to answer the extremely practical question of whether a signature is genuine or forged.”²⁶⁸ The court reasoned that since experts “see things in handwriting that laypeople do not—both because of analysts’ training in the minutiae of loops, swoops, and dotted ‘i’s, and because of the volume of handwriting they inspect,” such testimony is helpful to the jury. Like the Tenth Circuit in *Foust*, the Sixth Circuit recognized that “handwriting analysis may not boast the ‘empirical’ support underpinning scientific disciplines,” but ruled that it was properly admitted as a proper area of expertise based on specialized or technical expertise.²⁶⁹

Both *Foust* and *Mallory* downplayed the concerns raised about foundational reliability, choosing to cite older, pre-NRC and PCAST report cases, and to cast the evidence as technical or specialized rather than scientific. Both courts determined that as the specialty was not scientific, the traditionally applied *Daubert* factors were of less concern.²⁷⁰ In light of the 2023 amendments to Federal Rule of Evidence 702, as discussed earlier, courts may be required to more robustly address the concerns about foundational reliability that this specialty poses. These two cases also highlight potential questions courts may have about the need for proof of reliability as applied. Both courts appeared to rely on the fact that the experts were qualified and had years of experience. Indeed, many other courts have relied on experience and qualifications as proxies for reliability of FDE as applied.²⁷¹ There are two potential concerns with this approach. First, some commentators have argued that experience alone, at least in relatively subjective disciplines, is not a substitute for proficiency testing in establishing expert qualification. Second, the fact that an expert is proficient in a method does not establish that the method was reliably applied in a given case.²⁷² Reliability as applied could instead be established through internal validation, case file documentation, standards and protocols showing avoidance of contextual bias, the

267. *Foust*, 989 F.3d at 846–47.

268. *United States v. Mallory*, 902 F.3d 584, 593 (6th Cir. 2018).

269. *Id.*

270. *Id.* at 593–94.

271. See also Giannelli et al., Scientific Evidence, *supra* note 209, § 1.03[2] (noting the tendency of some courts to elide qualifications with reliability inquiries, particularly with experiential-knowledge-based specialties).

272. Garrett & Mitchell, *supra* note 19, at 909 (“Experts should be qualified based on empirical evidence of their proficiency before addressing whether their methods used and conclusions reached are valid and reliable.”).

nature of the samples examined, and the examiner's own testimony as to what they did and the conclusions they drew.

Other courts have been more amenable to hearing challenges to FDE. Although the excluded testimony at issue in *United States v. Johnsted* related to hand printing rather than hand writing, the court expressed doubts about the foundational reliability of the entire field of handwriting as well:

This lack of testing has serious repercussions on a practical level: because the entire premise of interpersonal individuality and intrapersonal variations of handwriting remains untested in reliable, double blind studies, the task of distinguishing a minor intrapersonal variation from a significant interpersonal difference—which is necessary for making an identification or exclusion—cannot be said to rest on scientifically valid principles. The lack of testing also calls into question the reliability of analysts' highly discretionary decisions as to whether some aspect of a questioned writing constitutes a difference or merely a variation; without any proof indicating that the distinction between the two is valid, those decisions do not appear based on a reliable methodology. With its underlying principles at best half-tested, handwriting analysis itself would appear to rest on a shaky foundation.²⁷³

In another federal district court case, *Almeciga v. Center for Investigative Reporting, Inc.*,²⁷⁴ the court excluded on *Daubert* grounds testimony about whether writing was real or disguised, opining that the field is relatively subjective and lacks adequate testing:

[T]here are no studies, to this Court's knowledge, that have evaluated the extent to which the angle at which one writes or the curvature of one's loops distinguish one person's handwriting from the next. Precisely what degree of variation falls within or outside an expected range of natural variation in one's handwriting—such that an examiner could distinguish in an objective way between variations that indicate different authorship and variations that do not—appears to be completely unknown and untested. Ditto the extent to which such a range is affected by the use of different writing instruments or the intentional disguise of one's natural hand or the passage of time. Such things could be tested and studied, but they have not been; and this by itself renders the field unscientific in nature.²⁷⁵

273. *United States v. Johnsted*, 30 F. Supp. 3d 814, 818 (W.D. Wis. 2013). *See also* *United States v. Fujii*, 152 F. Supp. 2d 939, 940 (N.D. Ill. 2000) (holding expert testimony concerning Japanese handprinting inadmissible; "Handwriting analysis does not stand up well under the *Daubert* standards. Despite its long history of use and acceptance, validation studies supporting its reliability are few, and the few that exist have been criticized for methodological flaws."); *United States v. Lewis*, 220 F. Supp. 2d 548 (S.D. W. Va. 2002); *United States v. Saelee*, 162 F. Supp. 2d 1097 (D. Alaska 2001); *Almeciga v. Ctr. for Investigative Reporting, Inc.*, 185 F. Supp. 3d 401 (S.D.N.Y. 2016).

274. 185 F. Supp. 3d 401 (S.D.N.Y. 2016).

275. *Id.* at 419.

The court also relied on what it described as a dearth of peer review and added that existing studies are “not sufficiently robust or objective to lend credence to the proposition that handwriting comparison is a scientific discipline.”²⁷⁶ The court stated that the known error rate is poor and even worse when there is a question of a disguised signature.²⁷⁷ Finally, the court discussed the lack of standards as to how many similarities are needed for a “match” or how many exemplars must be considered. In the court’s view, many conclusions by questioned document examiners are highly subjective, “bottom line” decisions whether there is a “match.” Summing up, the court wrote: “It remains the case that the methodology has not been subject to adequate testing or peer review, that error rates for the task at hand are unacceptably high, and that the field sorely lacks internal controls and standards, and so forth.”²⁷⁸ The court did not rule categorically that all handwriting comparison is inadmissible but noted the importance for caution in admitting testimony from an FDE, particularly on an opinion on authorship. More specifically, the court suggested that a trial judge “should not [admit questioned document testimony] without carefully evaluating whether the examiner has actual expertise in regard to the specific task at hand,” including proficiency testing.²⁷⁹ The court concluded that in this case, the evidence of the witness’s expertise was insufficient.²⁸⁰

Some district courts have endorsed a third view. These courts limit the reach of the examiner’s opinion, permitting expert testimony about similarities and dissimilarities between exemplars but not an ultimate conclusion that the defendant was the author (common authorship opinion) of the questioned document.²⁸¹ The expert is allowed to testify about “the specific similarities and

276. *Id.* at 420–21.

277. *Id.* at 422. Studies “suggest that while forensic document examiners might have some arguable expertise in distinguishing an authentic signature from a close forgery, they do not appear to have much, if any, facility for associating an author’s natural handwriting with his or her disguised handwriting.”

278. *Id.* at 424. *Accord Johnston*, 30 F. Supp. 3d 814.

279. *Almeciga*, 185 F. Supp. 3d at 424.

280. *But see* *United States v. Pitts*, No. 16–CR–550, 2018 WL 1116550 (E.D.N.Y. Feb. 26, 2018) (noting that many courts have admitted such evidence in the Second Circuit and distinguishing *Almeciga* as it involved a potentially forged signature, not just a comparison of known and unknown samples).

281. *See, e.g., United States v. Oskowitz*, 294 F. Supp. 2d 379, 384 (E.D.N.Y. 2003) (“Many other district courts have similarly permitted a handwriting expert to analyze a writing sample for the jury without permitting the expert to offer an opinion on the ultimate question of authorship.”); *United States v. Rutherford*, 104 F. Supp. 2d 1190, 1194 (D. Neb. 2000):

[T]he Court concludes that [the examiner]’s testimony meets the requirements of Rule 702 to the extent that he limits his testimony to identifying and explaining the similarities and dissimilarities between the known exemplars and the questioned documents. [The examiner] is precluded from rendering any ultimate conclusions on authorship of the questioned documents and is similarly precluded from testifying to the degree of confidence or certainty on which his opinions are based.

idiosyncrasies between the known writings and the questioned writings, as well as testimony regarding, for example, how frequently or infrequently in his experience, [the expert] has seen a particular idiosyncrasy.”²⁸² As the justification for this limitation, these courts often state that the examiners’ claimed ability to individuate lacks “empirical support.”²⁸³

As explained at the end of the section titled “Fingerprint Evidence” above, the 2023 amendment to Rule 702 may require courts to engage in more robust gatekeeping to evaluate both the foundational reliability and the reliability as applied of handwriting comparison evidence. The new amendment to Rule 702 may require a reconsideration of prior decisions focusing solely on weight and not admissibility of such evidence. As the advisory committee’s notes clarify, “many courts have held that the critical questions of the sufficiency of an expert’s basis, and the application of the expert’s methodology, are questions of weight and not admissibility. These rulings are an incorrect application of Rules 702 and 104(a).”²⁸⁴ As the notes recommend, “[j]udicial gatekeeping is essential.” As explained earlier, several gatekeeping options are available to the courts, along with the option of providing jury instructions to aid jurors in evaluating the evidence. With the enactment of the 2023 amendment to Rule 702, the judge must be satisfied that the proponent of the evidence has met each of the rule’s requirements by a preponderance. If the judge is not satisfied, she should exclude those opinions. Additionally, the judge may appropriately limit any opinion to exclude overstatement about the conclusions.

Firearms and Toolmark Analysis

Introduction

When ammunition travels through a firearm, the firearm leaves indentations or marks on the ammunition that arguably may be used to link it with a particular firearm or class of firearms. To that end, firearm and toolmark (sometimes referred to as FATM or FA/TM) examiners analyze fired (“spent”) ammunition found at crime scenes. They compare the marks on such ammunition either to the marks on other ammunition, to determine if both may have been fired from the same firearm or same category of firearm, or to a test fire from a particular suspected weapon, to determine if the ammunition might have been fired from

See also United States v. Hines, 55 F. Supp. 2d 62, 69 (D. Mass. 1999) (expert testimony concerning the general similarities and differences between a defendant’s handwriting exemplar and a stick-up note was admissible while the specific conclusion that the defendant was the author was not).

282. United States v. Van Wyk, 83 F. Supp. 2d 515, 524 (D.N.J. 2000).

283. United States v. Hidalgo, 229 F. Supp. 2d 961, 967 (D. Ariz. 2002).

284. Fed. R. Evid. 702 advisory committee’s note to 2023 amendment.

that weapon or a similar class of weapon. A related technique is used to compare the marks left on a surface of interest (like at a crime scene) to the marks left by a known tool (like a screwdriver) to determine if the known tool might have left the mark in question.

For decades, firearms and toolmark identification evidence has been widely accepted by courts.²⁸⁵ As with other feature comparison evidence, it was accepted with little critical scrutiny of its reliability and accuracy. More recently, the technique has been increasingly challenged. As with many disciplines, it is least controversial when used for excluding firearms, rather than identifying an individual firearm—exclusion as opposed to inclusion.

Concerns about the accuracy of firearm/toolmark analysis when used for source attribution are not merely hypothetical. The National Registry of Exonerations documents wrongful convictions based on firearms and toolmark analysis.²⁸⁶ Most notable is likely the wrongful conviction of Anthony Ray Hinton, who spent thirty years on death row for a murder he did not commit based on a firearm misidentification, before the U.S. Supreme Court reversed his conviction.²⁸⁷ Firearms misidentifications have also resulted in crime lab audits and lost accreditations.²⁸⁸

285. See *Gardner v. United States*, 140 A.3d 1172, 1183 (D.C. 2016) (discussing how District of Columbia courts have “allowed the admission of expert testimony concerning ballistics comparison matching techniques” for “decades”) (citing *Laney v. United States*, 294 F. 412, 416 (D.C. Cir. 1923)).

286. See, e.g., National Registry of Exonerations, *Patrick Pursley*, <https://perma.cc/Q4VG-E5JC> (last updated 4/29/2023); *People v. Pursley*, 2018 IL App (2d) 170227–U (2018). See also Brandon Garrett, *Siggers’ Firearms Exoneration*, Duke L. Forensic Forum (Oct. 23, 2018), <https://perma.cc/2HHS-HUYC> (wrongful conviction of Darrell Siggers); Craig Cooley & Gabriel Oberfield, *Increasing Forensic Evidence’s Reliability and Minimizing Wrongful Convictions: Applying Daubert Isn’t the Only Problem*, 43 Tulsa L. Rev. 285, 337–38 (2007) (wrongful conviction of Charles Stielow); Simon A. Cole, *Implicit Testing: Can Casework Validate Forensic Techniques?*, 46 *Jurimetrics J.* 117, 126–27 (2006). The Association of Firearm and Tool Mark Examiners (AFTE) has their own journal to publish studies; for concessions from the AFTE journal, see Bruce Moran, *A Report on the AFTE Theory of Identification & Range of Conclusions for Tool Mark Identification & Resulting Approaches to Casework*, 34 *AFTE J.* 227 (2002) (“In the 1980s some striated toolmark misidentifications resulting from a poor understanding of toolmark criteria for identification were experienced. An increasing need to address problems of applying subjective criteria became apparent.”); Evan E. Hodge, *Guarding Against Error*, 20 *AFTE J.* 290 (1988) (noting that “most of us [firearms examiners] know someone who has committed serious error” and describing misidentification by another examiner of the wrong .45 caliber firearm because of cognitive bias and pressure from prosecutors); Lowell Bradford, *Forensic Firearms Identification: Competence or Incompetence*, 11(2) *AFTE J.* 12 (1979) (“[a]n appalling number of misidentifications have been found in the firearm identification field”).

287. See *Hinton v. Alabama*, 571 U.S. 263 (2014); National Registry of Exonerations, *Anthony Ray Hinton*, <https://perma.cc/YJ3J-UQRU> (last updated Aug. 1, 2017).

288. In 2008, the Michigan State Police Forensic Science Division conducted an audit of the Detroit Police Department’s firearms unit at the request of the Wayne County Prosecutor’s Office and the Detroit Police Department chief. The audit included a random reanalysis of 283 cases. In 10% of the reanalyzed cases, the firearms unit had made false identifications or false exclusions.

Basic terms. Typically, three types of firearms—rifles, handguns, and shotguns—are encountered in criminal investigations. Rifles and handguns are classified according to their caliber. The caliber is the diameter of the bore of the firearm; the caliber is expressed in either hundredths or thousandths of an inch (e.g., .22, .45, .357 caliber) or millimeters (e.g., 7.62 mm). The two major types of handguns are revolvers²⁸⁹ and semiautomatic pistols. A major difference between the two is that when a semiautomatic pistol is fired, the cartridge case is automatically ejected and, if recovered at the crime scene, could help link the case to the firearm from which it was fired. In contrast, when a revolver is discharged the case is not ejected. In terms of ammunition, rifle and handgun cartridges consist of the projectile (bullet),²⁹⁰ case,²⁹¹ propellant (powder), and primer. The primer contains a small amount of an explosive mixture, which detonates when struck by the firing pin. When the firing pin detonates the primer, an explosion occurs that ignites the propellant. The most common modern propellant is smokeless powder.

The barrels of modern rifles and handguns are rifled; that is, parallel spiral grooves are cut into the inner surface, or bore, of the barrel. The surfaces between the grooves are called lands. The lands and grooves twist in a direction: right twist or left twist. For each type of firearm produced, the manufacturer specifies the number of lands and grooves, the direction of twist, the angle of

Michigan State Police Forensic Science Division, *Audit of the Detroit Police Department Forensic Services Laboratory Firearms Unit* (2008); see also Nick Bunkley, *Detroit Police Lab is Closed After Audit Finds Serious Errors in Many Cases*, N.Y. Times, Sept. 25, 2008, <https://www.nytimes.com/2008/09/26/us/26detroit.html>. Similarly, a failed proficiency test by a firearms examiner in 2017 launched multiple audits of the D.C. Department of Forensic Sciences. The audits revealed that three firearms examiners had misidentified cartridge cases in ongoing cases. The D.C. Department of Forensic Sciences lost its accreditation in 2021, after which the laboratory fired all of its firearms personnel, and set to review every firearms case from the last decade. See Spencer S. Hsu & Keith L. Alexander, *Forensic Errors Trigger Reviews of D.C. Crime Lab Ballistics Unit, Prosecutors Say*, Wash. Post, Mar. 24, 2017, <https://perma.cc/28FH-LWTT>; Jack Moore, *Sweeping Report Urges DC to Review Every Case Handled by Firearms, Fingerprint Units at Troubled Crime Lab*, WTOP News, Dec. 14, 2021, <https://perma.cc/GFN8-BYQK>; Jack Moore, *Officials Now Expect DC Crime Lab to Remain Sidelined Until Next Spring*, WTOP News, Mar. 31, 2022, <https://perma.cc/6LMQ-P8RM>.

289. Revolvers have a cylindrical magazine that rotates behind the barrel. The cylinder typically holds five to nine cartridges, each within a separate chamber. When a revolver is fired, the cylinder rotates and the next chamber is aligned with the barrel. A single-action revolver requires the manual cocking of the hammer; in a double-action revolver the trigger cocks the hammer. See generally BulletPoints Project, *Guns 101*, <https://perma.cc/6A43-DYDT> (last visited Dec. 5, 2024) (state-funded firearms injury prevention site with basic information about firearms).

290. Bullets are generally composed of lead and small amounts of other elements (hardeners). They may be completely covered (jacketed) with another metal or partially covered (semi-jacketed). See *id.* (discussing types and basics of ammunition).

291. Cartridge cases are generally made of brass. *Id.* They are sometimes referred to as “casings” by courts and laypeople, but firearms examiners do not use that term.

twist (pitch), the depth of the grooves, and the width of the lands and grooves. As a bullet passes through the bore, the lands and grooves force the bullet to rotate, giving it stability in flight and thus increased accuracy. Shotguns are “smooth-bore” firearms; that is, they do not have lands and grooves.

The Firearms Comparison Method and Its Claims

Bullet identification involves a comparison of the evidence bullet and a test bullet fired from the firearm.²⁹² The two bullets are examined by means of a comparison microscope, which permits a split-screen view of the two bullets and manipulation in order to attempt to align the striations (marks) on the two bullets.

The theory behind this comparative analysis is that barrels are machined during the manufacturing process, and imperfections in the tools used in the machining process are imprinted on the bore. The subsequent use of the firearm could add further individual imperfections. For example, mechanical action (erosion) caused by the friction of bullets passing through the bore of the firearm could produce accidental imperfections. Similarly, chemical action (corrosion) caused by moisture (rust), as well as primer and propellant chemicals, could produce other imperfections. However, the prevalence of certain imperfections remains unknown.

When a bullet is fired, microscopic striations are imprinted on the bullet surface as it passes through the bore of the firearm. These bullet markings are produced by the imperfections in the bore. Because these imperfections are randomly produced, examiners assume that they are unique to each firearm. Currently, there is no statistical basis for this assumption.

The condition of a firearm or evidence bullet may preclude a conclusion. For example, there may be insufficient marks on the bullet or, because of mutilation, an insufficient amount of the bullet may have been recovered. Likewise, if the bore of the firearm has changed significantly as a result of erosion or corrosion, a conclusion may be impossible. (Unlike fingerprints, firearms change over time.) In these situations, the examiner may render a “no conclusion” determination.

Firearms comparisons are made based on examinations of so-called class, subclass, and individual characteristics of bullet or cartridge cases. The class characteristics of a firearm result from design factors prior to manufacture. They include caliber and rifling specifications: (1) the land and groove diameters,

292. Test bullets are obtained by firing a firearm into a recovery box or bullet trap, which is usually filled with cotton, or into a recovery tank, which is filled with water. *See generally* NIST, Forensic Science Program, *Firearms and Toolmarks*, <https://perma.cc/9ATM-CZNX> (last updated Dec. 1, 2024) (discussing basics of firearms testing).

(2) the direction of rifling (left or right twist), (3) the number of lands and grooves, (4) the width of the lands and grooves, and (5) the degree of the rifling twist. Generally, a .38-caliber bullet with six land and groove impressions and with a right twist could have been fired only from a firearm with these characteristics. Such a bullet could not have been fired from a .32-caliber firearm, or from a .38-caliber firearm with a different number of lands and grooves or a left twist. According to the Association of Firearm and Tool Mark Examiners (AFTE),²⁹³ subclass characteristics are discernible surface features that are more restrictive than class characteristics in that they are (1) “produced incidental to manufacture,” (2) “relate to a smaller group source (a subset to which they belong),” and (3) can arise from a source that changes over time.²⁹⁴ The AFTE states that “[c]aution should be exercised in distinguishing subclass characteristics from class characteristics.”²⁹⁵ Finally, according to the AFTE, imperfections in the machining process and subsequent use of a firearm can cause individual characteristics that are thought to be unique to a particular firearm. Experts have warned that “caution should be exercised in distinguishing subclass characteristics from individual characteristics.”²⁹⁶

In terms of examiners’ ultimate opinions, the FBI’s ULTR on firearms and toolmark examinations directs examiners to report one of three conclusions: (1) source identification; (2) source exclusion; or (3) inconclusive.²⁹⁷ AFTE, in contrast, advises examiners that a source conclusion can be made when “sufficient agreement” exists in the patterns of two sets of marks. Agreement is sufficient when it “exceeds the best agreement demonstrated between toolmarks known to have been produced by different tools and is consistent with agreement demonstrated by toolmarks known to have been produced by the same tool,” such that “the likelihood another tool could have made the mark is so remote as to be considered a practical impossibility.”²⁹⁸

Although a conclusion about what marks exist is in one sense based on relatively objective data (the striations on the bullet surface), the AFTE explains that

293. AFTE is the leading professional organization in the field. There is also the Organization of Scientific Areas (OSAC) Firearms and Toolmarks Subcommittee, which promulgates standards and guidelines for examiners.

294. *Theory of Identification*, Association of Firearm and Toolmark Examiners, 30 AFTE J. 86, 88 (1998).

295. *Id.*

296. Robert M. Thompson, Program Manager, Forensic Data Systems, Office of Law Enforcement Standards, NIST, *Firearm Identification in the Forensic Laboratory*, Nat’l Dist. Att’y’s Ass’n (2010), <https://perma.cc/P37Y-ASNM>.

297. See U.S. Dep’t of Justice, Uniform Language for Testimony and Reports for the Forensic Firearms/Toolmarks Discipline Pattern Examination (2023), <https://perma.cc/4N57-TDAG>.

298. Association of Firearm & Tool Mark Examiners, *AFTE Theory of Identification*, <https://perma.cc/VTS2-HER3> (last visited Dec. 5, 2024).

the examiner's ultimate conclusion as to whether a projectile was fired from a particular weapon is a largely subjective judgment. The AFTE describes the traditional pattern recognition methodology as "subjective in nature, founded on scientific principles and based on the examiner's training and experience."²⁹⁹ There are no objective criteria governing this determination: "Ultimately, unless other issues are involved, it remains for the examiner to determine for himself the modicum of proof necessary to arrive at a definitive opinion."³⁰⁰

Another firearms comparison technique is cartridge case identification. Cartridge case identification is based on the same theory of random markings as bullet identification.³⁰¹ As with barrels, the belief is that defects produced in manufacturing leave distinctive characteristics on the breech face, firing pin, chamber, extractor, and ejector. Later use of the firearm is believed to produce additional defects. When the trigger is pulled, the firing pin strikes the primer of the cartridge, causing the primer to detonate. This detonation ignites the propellant (powder). In the process of combustion, the powder is converted rapidly into gases. The pressure produced by this process propels the bullet from the weapon and also forces the "base" (or "head") of the cartridge case backward against the breech face, imprinting breech face marks on the base of the cartridge case.³⁰² In turn, cartridge case identification involves a comparison of the case recovered at the crime scene and a test case obtained from the firearm after it has been fired. Shotgun shell cases may be analyzed in this way, as well. As in bullet identification, the comparison microscope is used in the examination, and the examiner's findings are, according to AFTE, "subjective in nature, based on one's training and experience."³⁰³

These ammunition identification techniques, like in other disciplines, are being increasingly automated. The current imaging system for identifying

299. *Theory of Identification*, AFTE, *supra* note 294, at 86.

300. Joseph L. Peterson et al., *Crime Laboratory Proficiency Testing Research Program*, Nat'l Inst. L. Enforcement & Crim. Just. 207 (1978), <https://perma.cc/BDJ8-JRNP>.

301. Gerald Burrard, *The Identification of Firearms and Forensic Ballistics* 107 (1962). However, bullet and cartridge case identifications differ in several respects. Because the bullet is traveling through the barrel at the time it is imprinted with the bore imperfections, these marks are "sliding" imprints, called striated marks. In contrast, the cartridge case receives "static" imprints, called impressed marks. *Id.* at 145.

302. Ammunition itself also has "class" characteristics, including markings on the head of the cartridge case, known as "head stamps," as well as bullet characteristics, such as caliber, type (such as hollow-point), weight, and jacketing, which may link cartridge cases or expelled projectiles at a scene to others collected elsewhere (such as unexpended cartridges within a seized firearm or at a search location), by caliber, brand, or type. *See, e.g.*, Defense Intelligence Agency, *Small-Caliber Ammunition Identification Guide*, Vol. 2 (1985), <https://perma.cc/U4Q2-NXE9>.

303. Eliot Springer, *Toolmark Examinations—A Review of Its Development in the Literature*, 40 J. Forensic Scis. 964, 966–67 (1995), <https://doi.org/10.1520/JFS13864J>.

bullets is the Integrated Ballistics Information System (IBIS).³⁰⁴ Automated systems screen bullets and cartridge cases in the database for possible “matches.”³⁰⁵ IBIS identifies a number of candidate bullets, and the examiner makes a final comparison with a microscope.³⁰⁶ The examiner has discretion to reject all the candidates.

The Toolmark Comparison Method and Its Claims

Toolmark comparisons rest on essentially the same theory as firearms identifications.³⁰⁷ Toolmark comparisons rest on the assumption that tools, and the marks they leave on surfaces like wood or metal, have both class and individual characteristics.³⁰⁸ For toolmarks, “class” characteristics might include the size of the mark/tool or the type of tool. For example, it may be possible to identify a mark (impression) left on wood as having been produced by a hammer, a knife, or a screwdriver. “Individual” characteristics are thought to include, for example, accidental imperfections produced by the machining process and subsequent use. When the tool is used, these characteristics are sometimes imparted onto the surface of another object struck by the tool.

As with firearms identification testimony, toolmark identification testimony is based on the largely subjective judgment of the examiner, who determines whether a sufficient number of marks of sufficient similarity are present to permit an identification.³⁰⁹ There are no easily repeatable or reproducible steps involving objective criteria governing the determination of whether there is a “match.”

304. See Jan De Kinder et al., *Reference Ballistic Imaging Database Performance*, 140 *Forensic Sci. Int.* 1 207 (2004), <https://doi.org/10.1016/j.forsciint.2003.12.002>; Ruprecht Nennstiel & Joachim Rahm, *An Experience Report Regarding the Performance of the IBIS™ Correlator*, 51 *J. Forensic Scis.* 24 (2006), <https://doi.org/10.1111/j.1556-4029.2005.00003.x>.

305. Richard E. Tontarski & Robert M. Thompson, *Automated Firearms Evidence Comparison: A Forensic Tool for Firearms Identification—An Update*, 43 *J. Forensic Scis.* 641, 641 (1998), <https://doi.org/10.1520/JFS16196J>.

306. *Id.*

307. See Springer, *supra* note 303, at 964:

The identification is based . . . on a series of scratches, depressions, and other marks which the tool leaves on the object it comes into contact with. The combination of these various marks ha[s] been termed toolmarks and the claim is that every instrument can impart a mark individual to itself.

308. While there are no studies assessing the phenomena of subclass characteristics for tools, in principle there is no reason to believe nongun tools would not also have subclass characteristics to the same extent as firearms.

309. See Springer, *supra* note 303, at 966–67 (“According to the Association of Firearms and Toolmarks Examiners’ Criteria for Identification Committee, interpretation of toolmark individualization and identification is still considered to be subjective in nature, based on one’s training and experience.”).

Scientific Assessments, Critiques, and Debates

The following discussion is focused on critiques of firearms analysis,³¹⁰ given that these will frame the litigation judges will face over admission of firearm comparison testimony.

Critics of firearm identification raise three primary challenges. First, they argue that it is premised on an empirically untested assumption: that every firearm leaves unique, individualized markings on spent ammunition, a direct trail back to the firearm in question. Without variability data to back up this assumption, critics argue that claims of source attribution are not yet sustainable. Second, they argue that the method of comparison itself is overly subjective, based largely on examiner judgment and experience and thus prone to cognitive bias. Third, critics argue that the method is insufficiently validated through well-designed studies as an accurate means of establishing whether ammunition was fired from a particular firearm. In particular, they argue that existing studies suffer serious design flaws that underestimate error rate. Linking ammunition to a larger class of firearms is less controversial, though also sometimes challenged, as discussed below.

The Crime Laboratory Proficiency Testing Program, begun in 1978, was the first published test of the accuracy of firearms analysis.³¹¹ In one test, 5.3% of the participating laboratories misidentified firearms evidence, and in another test 13.6% erred. These tests involved bullet and cartridge case comparisons. The Project Advisory Committee considered these errors “particularly grave in nature” and concluded that they probably resulted from carelessness, inexperience, or inadequate supervision.³¹² A third test required the examination of two bullets and two cartridge cases to identify the “most probable weapon” from which each was fired. The error rate was 28.2%. In later tests,

[e]xaminers generally did very well in making the comparisons. For all fifteen tests combined, examiners made a total of 2106 [bullet and cartridge case] comparisons and provided responses which agreed with the manufacturer responses 88% of the time, disagreed in only 1.4% of responses, and reported inconclusive results in 10% of cases.³¹³

310. This section focuses on firearms identification, although it notes where studies relate to toolmarks as well. The broader issues (i.e., subjectivity, lack of variability data, lack of well-designed validation studies) are likely to arise in toolmark litigation as well.

311. Peterson et al., *supra* note 300.

312. *Id.* at 207–08.

313. Joseph L. Peterson & Penelope N. Markham, *Crime Laboratory Proficiency Testing Results, 1978–1991, Part II: Resolving Questions of Common Origin*, 40 J. Forensic Scis. 1009, 1018 (1995), <https://doi.org/10.1520/JFS13871J>. The authors also stated:

The performance of laboratories in the firearms tests was comparable to that of the earlier LEAA study, although the rate of successful identifications actually was slightly over—88% vs. 91%.

In this same report, toolmark identification had higher error rates than firearm identification.³¹⁴

Citing this and other studies, the National Academy of Sciences in a 2008 Report on Ballistic Imaging concluded that “[t]he validity of the fundamental assumptions of uniqueness and reproducibility of firearms-related toolmarks has not yet been fully demonstrated.”³¹⁵ Specifically, it found that “[m]ost of th[e] existing firearms] studies are limited in scale and have been conducted by fire-arms examiners (and examiners in training) in state and local law enforcement laboratories as adjuncts to their regular casework.”³¹⁶ As psychologists have later noted, the insular nature of tests makes the error rate difficult to ascertain from such tests alone.³¹⁷ In addition, noting that firearms examiners often testify to a definite identification, the 2008 report cautioned, “examiners tend to cast their assessments in bold absolutes, commonly asserting that a ‘match’ can be made ‘to the exclusion of all other firearms in the world.’ Such comments cloak an inherently subjective assessment of a ‘match’ with an extreme probability statement that has no firm grounding and unrealistically implies an error rate of zero.”³¹⁸

The 2009 NRC Report likewise reviewed the existing validation studies, concluding that the technique can narrow down to a class of guns or tools, but was not yet sufficiently validated for determining that a particular gun or tool is the source of a mark or pattern:

Because not enough is known about the variabilities among individual tools and guns, we are not able to specify how many points of similarity are necessary for a given level of confidence in the result. Sufficient studies have not been done to understand the reliability and repeatability of the methods. The committee agrees that class characteristics are helpful in narrowing the pool of tools that may have left a distinctive mark. Individual patterns from manufacture or from wear might, in some cases, be distinctive enough to suggest one

Laboratories cut the rate of errant identifications by half (3% to 1.4%) but the rate of inconclusive responses doubled, from 5% to 10%.

Id. at 1019.

314. *Id.* at 1025 (“Overall, laboratories performed not as well on the toolmark tests as they did on the firearms tests.”).

315. National Research Council, *Ballistic Imaging* 81 (2008), <https://doi.org/10.17226/12162>.

316. *Id.* at 70. See also William A. Tobin, H. David Sheets & Clifford Spiegelman, *Absence of Statistical and Scientific Ethos: The Common Denominator in Deficient Forensic Practices*, 4 *Statistics & Pub. Pol’y* 1, 1 (2017), <https://doi.org/10.1080/2330443X.2016.1270175> (stating that the majority of firearms testing that has been done has been developed by firearms/toolmark examiners, an “insular communit[y] of nonscientist practitioners . . . who did not incorporate effective statistical methods”).

317. See Itiel E. Dror & Nicholas Scurich, *(Mis)use of Scientific Measurements in Forensic Science*, 2 *Forensic Sci. Int’l: Synergy* 333, 333 (2020), <https://doi.org/10.1016/j.fsisyn.2020.08.006>.

318. National Research Council, *supra* note 315, at 82.

particular source, but additional studies should be performed to make the process of individualization more precise and repeatable.³¹⁹

The report also identified a “fundamental problem with toolmark and firearms analysis” as being “the lack of a precisely defined process” or “specific protocol,” which the report said led to a lack of “well-characterized confidence limits” in the examiners’ conclusions.³²⁰

The 2016 PCAST Report has offered the most extensive analysis to date about *why* existing studies might underestimate error rate. The report explained that all but two existing firearms studies are either within-set studies or closed-set studies. In within-set studies, examiners analyze bullets to determine which came from the same firearm. The comparisons are not independent of each other (e.g., once bullet A matches bullet B, then if bullet C matches bullet A it must also match bullet B), which the report explained might artificially reduce the false-positive rate. In closed-set studies, examiners are asked to link each projectile to a gun in the set; all the right answers are included (all source guns are present). The problem with a closed-set study, according to the PCAST report, is that it is easier than real casework, for the same reason a multiple-choice question is easier (the right answer is definitely one of the choices given).³²¹

PCAST identified two existing firearms studies—the “Miami Dade study” and the “Ames study” (by the Ames Laboratory, a Department of Energy laboratory affiliated with Iowa State University)—that were not within-set or closed-set. The Miami Dade study was conducted by the Miami Dade police crime laboratory and involved a “partial open-set” where at least some source guns were not among those present.³²² Among the conclusive determinations made, four were false positives, resulting in a false positive rate of 1 in 48, with an upper bound (of the likely range of the true false positive rate) of 1 in 19. If one includes inconclusive determinations in the denominator (which PCAST argued was inappropriate), the false positive rate is lower.³²³ The Ames study was modeled after the FBI’s latent print study, and involved true open sets, where the examinations were fully independent of each other and where there was no indication whether the source gun was present. In the Ames study, there were 22 false positives and 735 “inconclusives” out of 2,158 total “different source” examinations where the ground truth (the correct call, that is) was “exclusion.”

319. 2009 NRC Report, *supra* note 37, at 154.

320. *Id.* at 155.

321. 2016 PCAST Report, *supra* note 20, at 107–09. The report notes that even where examiners are not explicitly told the study is a closed set, examiners can easily intuit this from the study design and answer sheet. *Id.* at 108.

322. *Id.* at 109. In the Miami Dade study, 165 examiners “were asked to assign a collection of 15 questioned samples, fired from 10 pistols to a collection of known standards; two of the 15 questioned samples came from a gun for which known standards were not provided. For these two samples, there were 188 eliminations, 138 inconclusives and 4 false positives.” *Id.*

323. *Id.*

PCAST calculated the false positive rate from this information as being 22 out of 1,443 (the number of conclusive examinations), leading to an estimated error rate as high as 2.2% or 1 in 46.³²⁴

Ultimately, PCAST concluded that only the Ames study was sufficiently well designed to support a finding of foundational validity of the method, in that they involved designers who were independent of law enforcement (even though the participants themselves were self-selected)³²⁵ and involved at least a partial open set. PCAST concluded that because of design flaws, other studies “seriously underestimate the false positive rate.” With only one well-designed study to back it, firearms analysis in PCAST’s view “still falls short of the scientific criteria for foundational validity.”³²⁶ PCAST urged that, if judges admitted firearms examination results in court, the testimony be accompanied by estimates of the false positive rate, based on the number of false positives in the examiners’ total number of conclusive examinations.³²⁷ The National District Attorneys Association disagreed with this conclusion in its response to PCAST, arguing that “the PCAST members are neither forensic firearm scientists performing casework nor did they participate as examiners in these validation studies” and that PCAST’s criteria for a well-designed study was “arbitrarily defined.”³²⁸

As judges might glean from the discussion above, there is a continuing debate about how to deal with “inconclusive” determinations in calculating a false positive rate. There are two separate issues. The first issue is whether to calculate a false positive rate based on the total number of examinations (including those where the examiner’s opinion was “inconclusive”) or solely based on conclusive examinations. The 2016 PCAST Report urged the latter approach, offering a hypothetical: “[C]onsider an extreme case in which a method had been tested 1000 times and found to yield 990 inconclusive results, 10 false positives, and no correct results. It would be misleading to report that the false positive rate was 1 percent (10/1,000 examinations). Rather, one should report that 100 percent of the conclusive results were false positives (10/10 examinations).”³²⁹

The second issue is whether an “inconclusive” determination should be treated as an error when the ground truth is an “exclusion”—for example, where a bullet was definitively not fired from Gun A but the examiner fails to call this an exclusion. In one sense, this call is not the same as a false identification because the examiner has not stated that the bullet was fired from the gun. Moreover, in real casework, “inconclusive” might be the most accurate inference from the

324. *Id.* at 110.

325. Alan H. Dorfman & Richard Valliant, *Inconclusives, Errors, and Error Rates in Forensic Firearms Analysis: Three Statistical Perspectives*, 5 *Forensic Sci. Int'l: Synergy* 1, 5 (2022), <https://doi.org/10.1016/j.fsisy.2022.100273> (noting self-selecting volunteers for Ames study).

326. 2016 PCAST Report, *supra* note 20, at 111–12.

327. *Id.*

328. Nat’l Dist. Att’y’s Ass’n, *supra* note 26, at 6.

329. 2016 PCAST Report, *supra* note 20, at 51–52.

evidence, given the conditions. In another sense, this call is erroneous in that it fails to exclude what is in fact an exclusion. This failure is especially concerning in a closed-set study where inconclusives are inherently a wrong response (because the source gun is always present), but is potentially concerning in nearly all validation studies, given that existing studies (unlike real casework) are designed to offer examiners sufficient information to correctly call exclusions when they are indeed exclusions.³³⁰ As of this writing, “[r]esearchers have yet to agree on whether and how inconclusives should be used when assessing examiner performance.”³³¹ The Food and Drug Administration, for its part, directs researchers facing this issue to calculate two error rates: one treating all inconclusives as a “positive” (here, an identification) and one treating inconclusives as a “negative” (here, an exclusion).³³² Judges should be aware of the debate and why “[i]t is clear that a well-founded characterization of inconclusives is critical for assessing the size of error rates estimated from the forensic firearms studies.”³³³

Since the 2016 PCAST Report, several other “open-set” validation studies of firearms comparison have been published, some by the AFTE in its own journal³³⁴ and others in peer-reviewed journals like the *Journal of Forensic Sciences*.³³⁵ Some of these new studies have also been critiqued on grounds similar to

330. See generally Itiel E. Dror & Glenn Langenburg, “Cannot Decide”: *The Fine Line Between Appropriate Inconclusive Determinations Versus Unjustifiably Deciding Not to Decide*, 64 J. Forensic Scis. 10, 11 (2018), <https://doi.org/10.1111/1556-4029.13854> (arguing that inconclusives should be treated as error where ground truth is exclusion); Maneka Sinha & Richard Gutierrez, *Signal Detection Theory Fails to Account for Real-World Consequences of Inconclusive Decisions*, 21 L., Probability & Risk 131 (2023), <https://doi.org/10.1093/lpr/mgad001> (same); Dror & Scurich, *supra* note 317, at 334 (same); Hal Arkes & Jonathan Jay Koehler, *Inconclusives and Error Rates in Forensic Science: A Signal Detection Theory Approach*, 20 L., Probability & Risk 153 (2021), <https://doi.org/10.1093/lpr/mgac005> (arguing that whether inconclusives are error depends on the context, given that “inconclusive” is not a state of nature); Dorfman & Valliant, *supra* note 325, at 6–7 (noting that inconclusives should at least sometimes be treated as errors); cf. Andrew Smith & Gary Wells, *Telling Us Less than What They Know: Expert Inconclusive Reports Conceal Exculpatory Evidence in Forensic Cartridge-Case Comparisons*, 13 J. Applied Res. Memory & Cognition 147 (2024), <https://doi.org/10.1037/mac0000138>.

331. Arkes & Koehler, *supra* note 330.

332. See U.S. Food & Drug Admin., Guidance for Industry and FDA Staff: Statistical Guidance on Reporting Results from Studies Evaluating Diagnostic Tests 18 (2007), <https://www.fda.gov/media/71147/download>.

333. Dorfman & Valliant, *supra* note 325, at 2.

334. See, e.g., Brandon A. Best & Elizabeth A. Gardner, *An Assessment of the Foundational Validity of Firearms Identification Using Ten Consecutively Button-Rifled Barrels*, 54 AFTE J. 28 (2022); Mark A. Keisler et al., *Isolated Pairs Research Study*, 50 AFTE J. 56 (2018).

335. See, e.g., Eric F. Law & Keith B. Morris, *Evaluating Firearm Examiner Conclusion Variability Using Cartridge Case Reproductions*, 66 J. Forensic Scis. 1704 (2021), <https://doi.org/10.1111/1556-4029.14758>; Keith L. Monson, Eric D. Smith & Eugene M. Peters, *Accuracy of Comparison Decisions by Forensic Firearms Examiners*, 68 J. Forensic Scis. 86 (2023), <https://doi.org/10.1111/1556-4029.15152>; Jaimie A. Smith, *Beretta Barrel Fired Bullet Validation Study*, 66 J. Forensic Scis. 547 (2021), <https://doi.org/10.1111/1556-4029.14604>. See also Max Guyll et al., *Validity of Forensic*

PCAST’s critiques of the Ames study,³³⁶ or because of their treatment of inconclusives³³⁷ or other issues related to “missingness” of data.³³⁸ Judges should be aware both that new studies are being published and of the recurring critiques of study design and error rate calculation that might be at the heart of the debate over such studies’ probative value in determining reliability.

Reliability as applied. As with fingerprints and handwriting analysis, evidence of the reliability as applied of firearms and toolmark comparison will typically be a mix of proficiency testing (which is also related to expert qualification), internal validation of any software or equipment used, documentation of whether the examiner followed standards and protocols, documentation of steps taken to avoid contextual and other bias, and evidence about the condition of the particular sample being examined. As with other methods, a judge might find based on the evidence presented that firearms and toolmark analysis is reliable as applied to offering one type of conclusion or answering one type of question, but not another.

In terms of proficiency testing, one study cited by proponents of firearms comparison testimony suggests a low 1.4% error rate based on a 1978–1991 study of results of proficiency tests given by the main private forensic proficiency testing firm, Collaborative Testing Services (CTS).³³⁹ In a 2005 CTS cartridge case examination, none of the 255 test-takers nationwide answered incorrectly.³⁴⁰ As one district court judge has noted, “[o]ne could read these results to mean that

Cartridge-Case Comparisons, 120 Proc. Nat’l Acad. Scis. e2210428120 (2023), <https://doi.org/10.1073/pnas.2210428120>.

336. After the 2016 PCAST Report, the Ames Laboratory and FBI collaborated on a second study, known as Ames II, originally available online as Stanley J. Bajic et al., Ames Laboratory, U.S. Dep’t of Energy, Technical Report No. ISTR-5220, *Report: Validation Study of the Accuracy, Repeatability, and Reproducibility of Firearms Comparison* (Oct. 7, 2020). The published version of the study in the *Journal of Forensic Sciences*, however, includes only FBI authors. See Keith L. Monson et al., *Repeatability and Reproducibility of Comparison Decisions by Firearms Examiners*, 68 J. Forensic Scis. 1721 (2023), <https://doi.org/10.1111/1556-4029.15318>. The FBI authors now state, without further explanation: “We acknowledge the contracted research contributions of Ames Laboratory personnel, who have declined authorship and individual acknowledgment.” *Id.* at 1738.

337. See, e.g., Dror & Scurich, *supra* note 317, at 334 (discussing Keisler et al., *supra* note 334); Michael Rosenblum et al., *Commentary on Guyll et al. (2023): Misuse of Statistical Method Results in Highly Biased Interpretation of Forensic Evidence 1* (Sept. 11, 2023), <https://perma.cc/VB9J-DV2C> (arguing that Guyll et al., *supra* note 335, “make a serious statistical error that leads to highly inflated claims about the probability that a cartridge case from a crime scene was fired from a reference gun”).

338. See, e.g., Khan & Carriquiry, *supra* note 102, at 4–5 (noting “missingness problems” affecting forensic error rate calculation, including self-selected participants, high rates of attrition, nonresponse, how inconclusive responses are used, low reproducibility (inter-examiner consistency) and repeatability (intra-examiner consistency)).

339. See, e.g., *United States v. Monteiro*, 407 F. Supp. 2d 351, 367 (D. Mass. 2006) (noting that government was citing and relying on Richard Grzybowski et al., *Firearm/Toolmark Identification: Passing the Reliability Test Under Federal and State Evidentiary Standards*, 35 AFTE J. 209, 213 (2003))

340. *Monteiro*, 407 F. Supp. 2d at 367.

the technique is foolproof, but the results might instead indicate that the test was somewhat elementary.”³⁴¹ Indeed, the CTS president has acknowledged that their customers prefer that tests are designed to be “easy.”³⁴² Some laboratories, like the Houston Forensic Science Center, have moved to “blind” proficiency testing where examiners do not know they are being tested.³⁴³ This solution has also been suggested as a means of accounting for artificially inflated accuracy rates where inconclusives are treated as non-errors.³⁴⁴

Case Law Development

Before 2005, courts had routinely allowed testimony of experts not only about class characteristics—or whether a bullet could have been fired from a particular firearm, or a mark could have been made by a particular tool³⁴⁵—but also more categorical statements,³⁴⁶ other than where experts were attempting particularly novel comparisons.³⁴⁷ In 2005, a federal district court judge ruled for the first time that an expert could not testify that recovered cases came from a

341. *Id.*

342. 2016 PCAST Report, *supra* note 20, at 57 n.133 (quoting CTS president).

343. Garrett & Mitchell, *supra* note 19, at 923.

344. *See, e.g.,* Arkes & Koehler, *supra* note 330 (suggesting blind proficiency testing as a possible solution to the inconclusive problem); Dorfman & Valliant, *supra* note 325, at 6–7 (same).

345. *E.g.,* People v. Horning, 102 P.3d 228, 236 (Cal. 2004) (expert “opined that both bullets and the casing could have been fired from the same gun . . . ; because of their condition he could not say for sure”); Luttrell v. Commonwealth, 952 S.W.2d 216, 218 (Ky. 1997) (expert “testified only that the bullets which killed the victim could have been fired from Luttrell’s gun”); United States v. Murphy, 996 F.2d 94, 99 (5th Cir. 1993) (allowing testimony of FBI expert “that the tools such as the screwdriver associated with Murphy ‘could’ have made the marks on the ignitions but that he could not positively attribute the marks to the tools identified with Murphy”).

346. *See, e.g.,* United States v. Bowers, 534 F.2d 186, 193 (9th Cir. 1976) (concluding that toolmark identification “rests upon a scientific basis and is a reliable and generally accepted procedure”); United States v. Hicks, 389 F.3d 514, 526 (5th Cir. 2004) (ruling that “the matching of spent shell casings to the weapon that fired them has been a recognized method of ballistics testing in this circuit for decades”); United States v. Foster, 300 F. Supp. 2d 375, 377 n.1 (D. Md. 2004) (“Ballistics evidence has been accepted in criminal cases for many years. . . . In the years since *Daubert*, numerous cases have confirmed the reliability of ballistics identification.”); United States v. Santiago, 199 F. Supp. 2d 101, 111 (S.D.N.Y. 2002) (“The Court has not found a single case in this Circuit that would suggest that the entire field of ballistics identification is unreliable.”); Commonwealth v. Whitacre, 878 A.2d 96, 101 (Pa. Super. Ct. 2005) (“no abuse of discretion in the trial court’s decision to permit admission of the evidence regarding comparison of the two shell casings with the shotgun owned by Appellant”).

347. *See, e.g.,* Ramirez v. State, 810 So. 2d 836, 849–51 (Fla. 2001) (rejecting testimony matching knife to a cartilage wound); Sexton v. State, 93 S.W.3d 96, 101 (Tex. Crim. App. 2002) (rejecting testimony matching collected fired cartridge cases to cases from *unfired* bullets found in appellant’s apartment based solely on magazine marks).

specific weapon “to the exclusion of every other firearm in the world.”³⁴⁸ Other courts similarly began to disallow certain statements suggesting certainty or definitive source attribution.³⁴⁹ Other courts continued to allow statements such as a firearm identification with “practical certainty.”³⁵⁰

Several courts in the past decade have, for the first time, more significantly limited firearm comparison testimony to statements about class characteristics or statements that a gun “cannot be excluded” as being a source of a mark or

348. *United States v. Green*, 405 F. Supp. 2d 104 (D. Mass. 2005).

349. *See, e.g., Williams v. United States*, 210 A.3d 734, 742 (D.C. 2019) (concluding that “the empirical foundation does not currently exist to permit these examiners to opine with certainty that a specific bullet can be matched to a specific gun,” and that the trial court plainly erred in allowing the testimony); *United States v. Diaz*, No. CR 05-00167 WHA, 2007 WL 485967, at *1 (N.D. Cal. Feb. 12, 2007) (allowing examiner to testify that a match has been made to a “reasonable degree of certainty in the ballistics field” but not “to the exclusion of all other firearms in the world”); *United States v. Glynn*, 578 F. Supp. 2d 567, 568 (S.D.N.Y. 2008) (disallowing “reasonable degree of ballistic certainty” testimony but allowing examiner to say that gun was “more likely than not” the source); *Gardner v. United States*, 140 A.3d 1172, 1177 (D.C. 2016) (holding that “firearms and toolmark expert may not give an unqualified opinion, or testify with absolute or 100% certainty, that based on ballistics feature comparison matching a fatal shot was fired from one firearm, to the exclusion of all other firearms”); *United States v. Willock*, 696 F. Supp. 2d 536, 546, 549 (D. Md. 2010) (Holding, based on a comprehensive magistrate judge’s report, that examiner “Sgt. Ensor shall not opine that it is a ‘practical impossibility’ for a firearm to have fired the cartridges other than the common ‘unknown firearm’ to which Sgt. Ensor attributes the cartridges.” Thus, “Sgt. Ensor shall state his opinions and conclusions without any characterization as to the degree of certainty with which he holds them.”); *United States v. Taylor*, 663 F. Supp. 2d 1170, 1180 (D.N.M. 2009):

[B]ecause of the limitations on the reliability of firearms identification evidence discussed above, Mr. Nichols will not be permitted to testify that his methodology allows him to reach this conclusion as a matter of scientific certainty. Mr. Nichols also will not be allowed to testify that he can conclude that there is a match to the exclusion, either practical or absolute, of all other guns. He may only testify that, in his opinion, the bullet came from the suspect rifle to within a reasonable degree of certainty in the firearms examination field.

See also United States v. Monteiro, 407 F. Supp. 2d 351, 374 (D. Mass. 2006) (allowing ballistics expert to testify to a “reasonable degree of certainty” but not a “statistical certainty”).

350. *See, e.g., United States v. Williams*, 506 F.3d 151, 161–62 (2d Cir. 2007) (allowing testimony of a ballistics “match” but cautioning that the opinion should not “be taken as saying that any proffered ballistic expert should be routinely admitted”); *United States v. McCluskey*, CR. No. 10-2734 JCH, 2013 WL 12335325, at *9–10 (D.N.M. Feb. 7, 2013) (allowing “practical certainty” but disallowing “reasonable degree of ballistic certainty” as a nonsensical nonscientific term); *United States v. Natson*, 469 F. Supp. 2d 1253, 1261 (M.D. Ga. 2007) (allowing statement of “100% degree of certainty”); *Commonwealth v. Meeks*, Nos. 2002-10961, 2003-10575, 2006 WL 2819423, at *50 (Mass. Super. Ct. Sept. 28, 2006) (allowing opinion of a “match”); *United States v. Casey*, 928 F. Supp. 2d 397, 399–400 (D.P.R. 2013) (finding that although “defendant challenges [the] conclusion that [the examiner] is 100% certain,” the court “remains faithful to the long-standing tradition of allowing the unfettered testimony of qualified ballistics experts”); *State v. Davidson*, 509 S.W.3d 156, 205 (Tenn. 2016) (“it’s like a fingerprint”); *State v. Anderson*, 624 S.E.2d 393, 397–98 (N.C. Ct. App. 2006) (no abuse of discretion in admitting bullet identification evidence).

pattern. In the often-cited case of *United States v. Tibbs*, a D.C. trial judge conducted a multiday evidentiary hearing, ultimately restricting the firearms testimony to a statement that the gun “cannot be excluded as the source of the cartridge case found on the scene,” rather than identification, because “[a]ny statement by the expert involving more certainty regarding the relationship between a casing and a firearm would stray into territory not presently supported by reliable principles and methods.”³⁵¹ The court recognized the “threshold design issues” of existing firearms/toolmark studies, which “limit their utility.”³⁵² A leading scientific evidence treatise calls *Tibbs* the “a high water mark for genuinely engaging with the imperfect and limited state of the scientific research underlying firearms identification research,”³⁵³ and several other courts since 2019 have similarly limited firearms examiner testimony.³⁵⁴ Indeed,

351. See *United States v. Tibbs*, 2016-CF1-19431, 2019 WL 4359486, at *23 (D.C. Super. Ct. Sept. 5, 2019).

352. *Id.* at *14.

353. Modern Scientific Evidence, *supra* note 138, at § 34:5.

354. See, e.g., *United States v. Briscoe*, No. 20-CR-1777 MV, 2023 WL 8096886 (D.N.M. Nov. 21, 2023) (citing *Tibbs* and allowing testimony but declining to allow evidence of “consistent with,” “sufficient agreement,” “match,” but allowing class characteristics); *Abruquah v. State*, 296 A.3d 961, 996–98 (Md. 2023) (allowing only testimony that gun was “consistent or inconsistent” with markings); *United States v. Cloud*, 576 F. Supp. 3d 827, 845 (E.D. Wash. 2021):

[I]f [examiner] intends to go beyond testimony that merely notes the recovered cartridge casings could not be excluded as having been fired from the recovered firearm, the Court will inform the jury that: (1) only three studies that meet the minimum design standard have attempted to measure the accuracy of firearm/toolmark comparison and (2) these studies found false positive rates that could be as high as 1 in 46 in one study, 1 in 200 in the second study, and 1 in 67 in the third study, though this study has yet to be published and subjected to peer review.

See also *United States v. Shipp*, 422 F. Supp. 3d 762, 783 (E.D.N.Y. 2019) (ruling that ballistics expert would be permitted to testify only that toolmarks on recovered bullet fragment and shell casing were consistent with having been fired from the recovered firearm); *United States v. Adams*, 444 F. Supp. 3d 1248, 1267 (D. Or. 2020) (limiting testimony to shared class characteristics only, such as caliber, type of firing pin, “barrel with six lands/grooves and right twist”); *People v. Ross*, 129 N.Y.S.3d 629, 642 (Sup. Ct. 2020) (limiting testimony to class characteristics only, and that the gun “cannot be ruled out” as the source, but disallowing phrases like “consistent with” or “sufficient agreement”); *Illinois v. Rickey Winfield*, 15 CR 14066-01 (Cook Cty. Cir. Ct. Feb. 8, 2023) (excluding ballistics testimony, concluding “[t]here are no objective forensic based reasons that firearms identification evidence belongs in any category of forensic science” and that “the junk evidence” “fails” the *Frye* test for admissibility (*Frye v. United States*, 293 F. 1013 (D.C. Cir. 1923))); *State v. Ghigliotti*, No. 17-0200154-1, slip op. at 39 (N.J. Super. Ct. Law Div. Aug. 23, 2019):

[T]his Court finds that the current state of firearms identification through the use of tool marks in the view of the scientific community does not support permitting [the expert] to frame his opinion that the identification of the evidence bullet in this case was “positive” or a “match” in terms of absolutely certainty.

Cf. *United States v. Medley*, No. PWG 17-242 (D. Md. Apr. 24, 2018) (pre-*Tibbs* opinion ruling that expert may not testify to a “match” opinion).

at least one court has cited *Tibbs* in rejecting a *defense* ballistics expert.³⁵⁵ At the same time, however, other courts continued to uphold admission.³⁵⁶

As explained at the end of the sections titled “Fingerprint Evidence” and “Handwriting Evidence,” the 2023 amendment to Rule 702 may require courts to engage in more robust gatekeeping to evaluate whether the opponent has proven by a preponderance of the evidence both the foundational reliability and the reliability as applied of firearm and toolmark comparison expert testimony.

Bitemark Evidence

Introduction

Forensic odontology is the field engaged in “examining, interpreting, and presenting dental and oral evidence for investigative and legal purposes.”³⁵⁷ Forensic bitemark examiners, a subset of forensic odontologists, purport to be able to identify a mark as a human bitemark and then attribute it to a particular person’s teeth.³⁵⁸ Other forensic odontologists focus only on human identification—that is, identifying people by their teeth.

Courts previously admitted bitemark analysis without much scrutiny into its reliability. In more recent years, scientists have overwhelmingly concluded that bitemark evidence is not reliable.³⁵⁹ As discussed below, research has

355. See *United States v. Harris*, 491 F. Supp. 3d 414, 424–25 (E.D. Wis. 2020) (rejecting on *Daubert* grounds proffered defense expert testimony that mark on car was not from a bullet, citing *Tibbs*).

356. See, e.g., *United States v. Hunt*, 63 F.4th 1229, 1249, n.18 (10th Cir. 2023) (allowing testimony based on CMS (consecutive matching striae) and noting that *Tibbs* did not deal with that method); *United States v. Perry*, 35 F.4th 293, 329–31 & n.23 (5th Cir. 2022) (holding that trial court’s admission of ballistics testimony under *Daubert* was not an abuse of discretion, notwithstanding the defendant’s arguments as to lack of peer review and standardization); *United States v. Graham*, No. 4:23-CR-00006, 2024 WL 688256, at *16 (W.D. Va. Feb. 20, 2024) (denying *Daubert* challenge to ballistics expert, but requiring expert to testify consistent with FBI’s new ULTR); *United States v. Blackman*, No. 18-CR-00728, 2023 WL 3440384, at *5 n.5, *9 (N.D. Ill. May 12, 2023) (allowing source attribution testimony so long as expert does not imply absolute certainty or “to the exclusion of all other” guns) (citing *United States v. Green*, 405 F. Supp. 2d 104, 124 (D. Mass. 2005)); *United States v. Harris*, 502 F. Supp. 3d 28, 33 (D.D.C. 2020) (denying *Daubert* challenge to ballistics expert, but requiring expert to testify consistent with FBI’s new ULTR).

357. NIST, *Forensic Odontology Subcommittee*, <https://perma.cc/5L4U-HTG6> (last updated Nov. 3, 2023).

358. See E.H. Dinkel, Jr., *The Use of Bite Mark Evidence as an Investigative Aid*, 19 J. Forensic Scis. 535 (1974), <https://doi.org/10.1520/JFS10208J>.

359. NIST Bitemark Report, *supra* note 110. As the Texas Forensic Science Commission wrote, “if anyone should take responsibility for the current state of bite mark comparison, it is the very organization of practitioners that, due to its glacial pace, reticence to publish critical data, and willingness to allow overstatements of science to go unchecked for decades, is facing a barrage of well-founded criticism.” Tex. Forensic Sci. Comm’n, *Forensic Bitemark Comparison Complaint*

documented that (1) skin is an unstable medium upon which to expect consistent imprint results; (2) dental molds are frequently similar in nature and challenging to distinguish; and (3) bitemark experts themselves have increasing difficulty determining first whether a blemish is a bitemark, then whether it is a human or animal bitemark, and finally, whether the bitemark “matches” to a dental mold. The uniqueness of dentition has also never been proven. The OSAC Forensic Odontology Subcommittee itself makes clear on its website that it “does not develop standards on bitemark recognition, comparison, and identification.”³⁶⁰

Still, as of this writing, some prosecutors and defense attorneys continue to offer bitemark evidence, and some courts continue to admit it.

The Method and Its Claims

Human identification. Forensic odontologists identify who a deceased person might be from human remains by comparing the decedent’s teeth to dental records of a known person. The human adult set of teeth (dentition) offers many points of comparison for these purposes. A set of teeth consists of thirty-two teeth, each with five anatomic surfaces. Restorations, with varying shapes, sizes, and restorative materials, may offer numerous additional points of comparison. Moreover, the number of teeth, prostheses, decay, malposition, malrotation, peculiar shapes, root canal therapy, bone patterns, bite relationship, and oral pathology may also provide points of comparison.³⁶¹

Although the assumption that every person’s full dentition is unique remains empirically unproven,³⁶² “the identification of human remains by their dental characteristics is” nonetheless “well established.”³⁶³ The reliability of human identification through dental records stems from the finite number of candidates to identify and the availability of full dentition on both remains and records.³⁶⁴

Filed by National Innocence Project on behalf of Steven Mark Chaney—Final Report, 1, 17 (Apr. 12, 2016), <https://www.txcourts.gov/media/1454500/finalbitemarkreport.pdf>.

360. NIST, *supra* note 357.

361. The identification is made by comparing the decedent’s teeth with antemortem dental records. See generally Javier Ata-Ali & Fadi Ata-Ali, *Forensic Dentistry in Human Identification: A Review of the Literature*, 6 J. Clin. & Exp. Dent. e162 (2014), <https://doi.org/10.4317/jced.51387> (explaining points of comparison, process, and types of records relied on).

362. See 2009 NRC Report, *supra* note 37, at 175 (“The uniqueness of the human dentition has not been scientifically established.”); NIST Bitemark Report, *supra* note 110, at i (finding a “lack of support” for this “key premise”).

363. 2009 NRC Report, *supra* note 37, at 173. See also Satomi Mizuno et al., *Validity of Dental Findings for Identification by Postmortem Computed Tomography*, 341 Forensic Sci Int’l 111507 (2022), <https://doi.org/10.1016/j.forsciint.2022.111507>.

364. See Iain A. Pretty & David J. Sweet, *The Scientific Basis for Human Bitemark Analyses—A Critical Review*, 41 Sci. & Just. 85, 88 (2001), [https://doi.org/10.1016/S1355-0306\(01\)71859-X](https://doi.org/10.1016/S1355-0306(01)71859-X)

Bitemark identification. Bitemark identification, by contrast to human remains identification, relies on impressions and bruising of the skin and comparing those to a suspect's teeth. In contrast to human remains identification, the number of potential suspects may be large, the examiner may have only a small number of impressions to work with, the marks may change over time, and the medium on which the marks are observed (like skin) may be unstable and allow shrinkage and distortion.³⁶⁵ Moreover, all five anatomic surfaces are not engaged in biting; only the edges of the front teeth come into play. In sum, bitemark identification depends not only on the uniqueness of each person's dentition but also on "whether there is a [sufficient] representation of that uniqueness in the mark found on the skin or other inanimate object."³⁶⁶ This uniqueness has yet to be proven.³⁶⁷

Of the several methods of bitemark analysis that have been reported, all involve three steps: (1) registration of both the bitemark and the suspect's dentition, (2) comparison of the dentition and bitemark, and (3) evaluation of the points of similarity or dissimilarity. The comparison may be either direct or indirect. A model of the suspect's teeth is used in direct comparisons; the model is compared to life-size photographs of the bitemark. Transparent overlays made from the model are used in indirect comparisons.³⁶⁸ The ultimate opinion of a bitemark examiner regarding individuation is a relatively subjective one.³⁶⁹ For example, there is no accepted minimum number of points of identity required

("A distinction must be drawn from the ability of a forensic dentist to identify an individual from their dentition by using radiographs and dental records and the science of bitemark analysis.").

365. See 2009 NRC Report, *supra* note 37, at 174–76 (noting these limitations).

366. Raymond D. Rawson et al., *Statistical Evidence for the Individuality of the Human Dentition*, 29 J. Forensic Scis. 245, 252 (1984), <https://doi.org/10.1520/JFS11656J>.

367. Mary A. Bush et al., *Statistical Evidence for the Similarity of the Human Dentition*, 56 J. Forensic Scis. 118, 118 (2010), <https://doi.org/10.1111/j.1556-4029.2010.01531.x> (observing significant correlations and nonuniform distributions of tooth positions as well as "matches" between dentitions and concluding that "statements of dental uniqueness with respect to bitemark analysis in an open population are unsupportable and that use of the product rule is inappropriate"); Mary A. Bush et al., *Similarity and Match Rates of the Human Dentition in Three Dimensions: Relevance to Bitemark Analysis*, 125 Int'l J. Legal Med. 779, 779 (2011), <https://doi.org/10.1007/s00414-010-0507-8> (explaining that three dimensional models reduce but do not limit random matches and "a zero match rate cannot be claimed for the population studied"); H. David Sheets et al., *Dental Shape Match Rates in Selected and Orthodontically Treated Populations in New York State: A Two-Dimensional Study*, 56 J. Forensic Scis. 621, 621 (2011), <https://doi.org/10.1111/j.1556-4029.2011.01731.x> (finding random dental shape "matches" and concluding that "statements of certainty concerning individualization in such populations should be approached with caution").

368. See David J. Sweet, *Human Bitemarks: Examination, Recovery, and Analysis*, in *Manual of Forensic Odontology* 162 (American Society of Forensic Odontology, 3d ed. 1997) ("The analytical protocol for bitemark comparison is made up of two broad categories. Firstly, the measurement of specific traits and features called a metric analysis, and secondly, the physical matching or comparison of the configuration and pattern of the injury called a pattern association.").

369. See Roland F. Kouble & Geoffrey T. Craig, *A Comparison Between Direct and Indirect Methods Available for Human Bite Mark Analysis*, 49 J. Forensic Scis. 111, 111 (2004), <https://doi.org/10.1520/JFS2001252> ("It is important to remember that computer-generated overlays still retain an

for a positive identification,³⁷⁰ and experts have testified to a wide range of points of similarity, from a low of eight points to a high of fifty-two.³⁷¹

In an attempt to develop a more objective method, in 1984 the American Board of Forensic Odontology (ABFO) promulgated guidelines for bitemark analysis, including a uniform scoring system.³⁷² According to the drafting committee, “[t]he scoring system . . . has demonstrated a method of evaluation that produced a high degree of reliability among observers.”³⁷³ Moreover, the committee characterized “[t]he scoring guide . . . [as] the beginning of a truly scientific approach to bite mark analysis.”³⁷⁴ In a subsequent letter, however, the drafting committee wrote:

While the Board’s published guidelines suggest use of the scoring system, the authors’ present recommendation is that all odontologists await the results of further research before relying on precise point counts in evidentiary proceedings. . . . [T]he authors believe that further research is needed regarding the quantification of bite mark evidence before precise point counts can be relied upon in court proceedings.³⁷⁵

Ultimately, the ABFO’s effort to develop a standardized bitemark methodology “failed . . . due to inter examiner discord and unreliable quantitative interpretation.”³⁷⁶ The ABFO has conceded that bitemark evidence cannot be used for individualization in “open population” cases, where there is an unknown number of potential suspects.³⁷⁷

element of subjectivity, as the selection of the biting edge profiles is reliant on the operator placing the ‘magic wand’ onto the areas to be highlighted within the digitized image.”).

370. See *Stubbs v. State*, 845 So. 2d 656, 669 (Miss. 2003) (“There is little consensus in the scientific community on the number of points which must match before any positive identification can be announced.”). Leigh Stubbs was exonerated in 2012. See Valena Beety, *Manifesting Justice: Wrongly Convicted Women Reclaim Their Rights* (2022).

371. E.g., *State v. Garrison*, 585 P.2d 563, 566 (Ariz. 1978) (10 points); *People v. Slone*, 143 Cal. Rptr. 61, 67 (Ct. App. 1978) (10 points); *People v. Milone*, 356 N.E.2d 1350, 1356 (Ill. App. Ct. 1976) (29 points); *State v. Sager*, 600 S.W.2d 541, 564 (Mo. Ct. App. 1980) (52 points); *State v. Green*, 290 S.E.2d 625, 630 (N.C. 1982) (14 points); *State v. Temple*, 273 S.E.2d 273, 279 (N.C. 1981) (8 points); *Kennedy v. State*, 640 P.2d 971, 976 (Okla. Crim. App. 1982) (40 points); *State v. Jones*, 259 S.E.2d 120, 125 (S.C. 1979) (37 points).

372. Am. Board of Forensic Odontology (ABFO), *Guidelines for Bite Mark Analysis*, 112 J. Am. Dental Ass’n 383 (1986).

373. Raymond D. Rawson et al., *Reliability of the Scoring System of the American Board of Forensic Odontology for Human Bite Marks*, 31 J. Forensic Scis. 1235, 1256 (1986), <https://doi.org/10.1520/JFS11903J>.

374. *Id.* at 1259.

375. Gerald L. Vale et al., *Letters to the Editor: Discussion of “Reliability of the Scoring System of the American Board of Forensic Odontology for Human Bite Marks,”* 33 J. Forensic Scis. 20 (1988).

376. C. Michael Bowers, *Problem-Based Analysis of Bitemark Misidentifications: The Role of DNA*, 159S Forensic Sci. Int’l S104, S106 (2006), <https://doi.org/10.1016/j.forsciint.2006.02.032>.

377. Am. Bd. of Forensic Odontology, Inc., *Diplomates Reference Manual* 102 (2015) (explaining that “[t]he ABFO does not support a conclusion of ‘The Biter’ in an open population case(s)”).

Scientific Assessments, Critiques, and Debates

Bitemark analysis is rarely used in the United States because of serious concerns about its reliability, in part inspired by high-profile DNA exonerations involving people convicted based on bitemark testimony. For example, in *State v. Krone*,³⁷⁸ two experienced experts concluded that the defendant had made the bitemark found on a murder victim. The defendant, however, was later exonerated through DNA testing.³⁷⁹ In *Otero v. Warnick*,³⁸⁰ a forensic dentist testified that the

plaintiff was the only person in the world who could have inflicted the bite marks on [the murder victim's] body. On January 30, 1995, the Detroit Police Crime Laboratory released a supplemental report that concluded that plaintiff was excluded as a possible source of DNA obtained from vaginal and rectal swabs taken from [the victim's] body.³⁸¹

In *Burke v. Town of Walpole*,³⁸² the expert concluded that “Burke’s teeth matched the bite mark on the victim’s left breast to a ‘reasonable degree of scientific certainty.’ That same morning . . . DNA analysis showed that Burke was excluded as the source of male DNA found in the bite mark on the victim’s left breast.”³⁸³ These are but a few of the exonerations involving false bitemark evidence.

The 2009 NRC Report concluded that neither “[t]he uniqueness of human dentition,” “[t]he ability of the dentition, if unique, to transfer a unique pattern to human skin and the ability of the skin to maintain that uniqueness” nor any “standard for the type, quality, and number of individual characteristics required to indicate that a bite mark has reached a threshold of evidentiary value” has been scientifically established.³⁸⁴ Moreover, as the report noted, “[t]here is no science on the reproducibility of the different methods of analysis that lead to conclusions about the probability of a match. This includes reproducibility between experts and with the same expert over time. Even when using the

378. 897 P.2d 621, 622, 623 (Ariz. 1995) (“The bite marks were crucial to the State’s case because there was very little other evidence to suggest Krone’s guilt.”; “Another State dental expert, Dr. John Piakis, also said that Krone made the bite marks . . . Dr. Rawson himself said that Krone made the bite marks. . .”).

379. See Mark Hansen, *The Uncertain Science of Evidence*, A.B.A. J., July 28, 2005, at 49, <https://perma.cc/G674-XKPB> (discussing *Krone*).

380. 614 N.W.2d 177 (Mich. Ct. App. 2000).

381. *Id.* at 178.

382. 405 F.3d 66 (1st Cir. 2005).

383. *Id.* at 73. See also Bowers, *supra* note 376, at S106–07 (citing several cases involving bitemarks and DNA exonerations); Mark Hansen, *Out of the Blue*, A.B.A. J., Feb. 1, 1996, at 50, 51, <https://perma.cc/GV32-SAC4> (DNA analysis of skin taken from fingernail scrapings of the victim conclusively excluded suspect).

384. 2009 NRC Report, *supra* note 37, at 175–76. See also *id.* at 176 (“Although the majority of forensic odontologists are satisfied that bite marks can demonstrate sufficient detail for positive identification, no scientific studies support this assessment, and no large population studies have been conducted.”).

[American Board of Forensic Odontology] guidelines, different experts provide widely differing results and a high percentage of false positive matches of bite marks using controlled comparison studies.”³⁸⁵ As a result, as early as 2009, there was already “continuing dispute over the value and scientific validity of comparing and identifying bite marks.”³⁸⁶ Indeed, because of the dispute, some odontologists well before the 2009 NRC Report had argued that “bitemark evidence should only be used to exclude a suspect.”³⁸⁷

In 2015, the ABFO conducted an internal study titled *Construct Validity of Bitemark Assessments Using the ABFO Bitemark Decision Tree*.³⁸⁸ This experiment presented 100 injuries to 39 ABFO board-certified forensic odontologists, to determine whether the injury at issue was a human bitemark and, if so, whether it had distinct identifiable arches and individual toothmarks.³⁸⁹ Thirty-nine forensic odontologists completed the survey and “came to unanimous agreement on just 4 of the 100 case studies. Of the initial 100, there remained just 8 case studies in which at least 90 percent of the analysts were still in agreement.”³⁹⁰

The Texas Forensic Science Commission, in a 2016 report, concluded that “there is no scientific basis for stating that a particular patterned injury can be associated to an individual’s dentition” and “[a]ny testimony describing human dentition as ‘like a fingerprint’ or incorporating similar analogies lacks scientific support.”³⁹¹ It ultimately urged that analyst testimony regarding the probability or weight of any association between a bitemark and an individual’s dentition has “no place in our criminal justice system because they lack any credible supporting data.”³⁹²

The 2016 PCAST Report likewise concluded that “available scientific evidence strongly suggests that examiners not only cannot identify the source of bitemark with reasonable accuracy, they cannot even consistently agree on whether an injury is a human bitemark.”³⁹³

Most recently, in 2023, NIST released a report, *Bitemark Analysis: A NIST Scientific Foundation Review*, reviewing bitemark analysis. The report found

forensic bitemark analysis lacks a sufficient scientific foundation because the three key premises of the field are not supported by the data. First, human

385. *Id.* at 174.

386. *Id.* at 173.

387. Iain A. Pretty, *A Web-Based Survey of Odontologist’s [sic] Opinions Concerning Bitemark Analyses*, 48 J. Forensic Scis. 1117, 1120 (2003), <https://doi.org/10.1520/JFS2003017>.

388. Adam J. Freeman & Iain A. Pretty, *Construct Validity of Bitemark Assessments Using the ABFO Decision Tree* (2015), <https://perma.cc/D7FZ-JAJX>.

389. Radley Balko, *A Bite Mark Matching Advocacy Group Just Conducted a Study that Discredits Bite Mark Evidence*, Wash. Post, Apr. 8, 2015, <https://perma.cc/S2DL-ZREX>.

390. *Id.*

391. Tex. Forensic Sci. Comm’n, *supra* note 359, at 11–12.

392. *Id.* at 12.

393. 2016 PCAST Report, *supra* note 20, at 9 (emphasis in original).

anterior dental patterns have not been shown to be unique at the individual level. Second, those patterns are not accurately transferred to human skin consistently. Third, it has not been shown that defining characteristics of those patterns can be accurately analyzed to exclude or not exclude individuals as the source of a bitemark.³⁹⁴

Case Law Development

With respect to the use of dentition to identify human remains, courts have readily accepted the method as a means of establishing the identity of a homicide victim,³⁹⁵ with some cases dating back to the nineteenth century.³⁹⁶

With respect to bitemarks, *People v. Marx*³⁹⁷ emerged as the seminal case. The case “came to be read as a global warrant to admit bite mark identification evidence whenever a person displaying apparent credentials chose to testify to an identification.”³⁹⁸ The cases that closely followed and relied on *Marx* also went to great lengths to extoll the “superior trustworthiness of the scientific bite mark approach.”³⁹⁹ After *Marx*, bitemark evidence became widely accepted.⁴⁰⁰ By 1992, it had been introduced or noted in 193 reported cases and accepted as

394. NIST Bitemark Report, *supra* note 110, at 24.

395. *E.g.*, *Wooley v. People*, 367 P.2d 903, 905 (Colo. 1961) (dentist compared his patient’s record with dentition of a corpse); *Martin v. State*, 636 N.E.2d 1268, 1272 (Ind. Ct. App. 1994) (dentist qualified to compare X-rays of one of his patients with skeletal remains of murder victim and make a positive identification); *Fields v. State*, 322 P.2d 431, 446 (Okla. Crim. App. 1958) (murder case in which victim was burned beyond recognition).

396. *See Commonwealth v. Webster*, 59 Mass. 295, 299–300 (1850) (remains of the incinerated victim, including charred teeth and parts of a denture, were identified by the victim’s dentist); *Lindsay v. People*, 63 N.Y. 143, 145–46 (1875).

397. 126 Cal. Rptr. 350 (Cal. Ct. App. 1975). The court in *Marx* avoided applying the *Frye* test, which requires acceptance of a novel technique by the scientific community as a prerequisite to admissibility. *See Frye v. United States*, 293 F. 1013 (D.C. Cir. 1923). According to the *Marx* court, the *Frye* test “finds its rational basis in the degree to which the trier of fact must accept, on faith, scientific hypotheses not capable of proof or disproof in court and not even generally accepted outside the courtroom.” 126 Cal. Rptr. at 355–56.

398. D. Michael Risinger, *Navigating Expert Reliability: Are Criminal Standards of Certainty Being Left on the Dock?*, 64 Alb. L. Rev. 99, 138 (2000).

399. *People v. Slone*, 143 Cal. Rptr. 61, 69 (Ct. App. 1978); *see also State v. Sager*, 600 S.W.2d 541, 569 (Mo. Ct. App. 1980) (characterizing bitemark evidence as “an exact science”).

400. Two Australian cases, however, excluded bitemark evidence. *See Lewis v The Queen* (1987), 29 A Crim R 267 (odontological evidence was improperly relied on, in that this method has not been scientifically accepted); *R v Carroll* (1985), 19 A Crim R 410 (“[T]he evidence given by the three odontologist is such that it would be unsafe or dangerous to allow a verdict based upon it to stand.”).

admissible in 35 states.⁴⁰¹ Some courts described bitemark comparison as an “exact science,”⁴⁰² and several cases took judicial notice of its validity.⁴⁰³ The first state supreme court to take judicial notice of the “general acceptance” of bitemark evidence was West Virginia in *State v. Armstrong*, prompting other state supreme courts to rule similarly.⁴⁰⁴ *State v. Armstrong*, however, relied on the findings in another bitemark identification case, that of Robert Lee Stinson in Wisconsin.⁴⁰⁵ Stinson was ultimately exonerated by DNA evidence in 2009.⁴⁰⁶ Hence, even the bitemark identification in the case undergirding what led to the national “general acceptance” of bitemark evidence was wrong.

Because law enforcement around the country has so significantly curtailed its reliance on bitemark evidence, there is little recent case law on the subject. Indeed, serious court evaluation of bitemarks has still been largely confined to civil postconviction habeas corpus cases and 42 U.S.C. § 1983 lawsuits for wrongful conviction and presentation of false evidence at trial.⁴⁰⁷ Still, some courts have continued to admit the evidence.⁴⁰⁸

401. Steven Weigler, *Bite Mark Evidence: Forensic Odontology and the Law*, 2 Health Matrix: J.L.-Med. 303, 303 (1992).

402. See *People v. Marsh*, 441 N.W.2d 33, 35 (Mich. Ct. App. 1989) (“the science of bite mark analysis has been extensively reviewed in other jurisdictions”); *Sager*, 600 S.W.2d at 569 (“an exact science”).

403. See *State v. Richards*, 804 P.2d 109, 112 (Ariz. Ct. App. 1990) (“[B]ite mark evidence is admissible without a preliminary determination of reliability. . . .”); *People v. Middleton*, 429 N.E.2d 100, 101 (N.Y. 1981) (“The reliability of bite mark evidence as a means of identification is sufficiently established in the scientific community to make such evidence admissible in a criminal case, without separately establishing scientific reliability in each case. . . .”); *State v. Armstrong*, 369 S.E.2d 870, 877 (W. Va. 1988) (judicially noticing the reliability of bitemark evidence).

404. *Armstrong*, 369 S.E.2d at 877.

405. *State v. Stinson*, 397 N.W.2d 136 (Wis. Ct. App. 1986).

406. See Jennifer D. Oliva & Valena E. Beety, *Discovering Forensic Fraud*, 112 Nw. U. L. Rev. 121 (2017).

407. See, e.g., *Keko v. Hingle*, 318 F.3d 639, 644 (5th Cir. 2003) (denying absolute immunity to forensic odontologist in § 1983 civil lawsuit following wrongful conviction); *Ege v. Yukins*, 380 F. Supp. 2d 852, 871 (E.D. Mich. 2005), *aff’d in part, rev’d in part on other grounds*, 485 F.3d 364 (6th Cir. 2007) (ruling “there is no question that the [bitemark] evidence in this case was unreliable and not worthy of consideration by a jury”); *In re Richards*, 371 P.3d 195, 211 (Cal. 2016) (court granting civil writ of habeas corpus ruling bitemark expert’s criminal trial testimony constituted material false evidence); *Stinson v. Milwaukee*, No. 09-C-1033, 2013 WL 5447916, at *12–13, *15–17 (E.D. Wis. Sept. 30, 2013), *aff’d in part, rev’d in part sub nom.* *Stinson v. Gauger*, 799 F.3d 833 (7th Cir. 2015) (denying absolute immunity to forensic odontologists in § 1983 civil lawsuit alleging fabrication and suppression of evidence).

408. See Aliza B. Kaplan & Janis C. Puracal, *It’s Not a Match: Why the Law Can’t Let Go of Junk Science*, 81 Alb. L. Rev. 895, 898 (2018) (notes courts that continue to admit bitemarks, relying on precedent); Michael J. Saks et al., *Forensic Bitemark Identification: Weak Foundations, Exaggerated Claims*, 3 J.L. & Biosciences 538, 546–47 (2016), <https://doi.org/10.1093/jlb/lsw045> (same).

Facial Recognition Software and Other Methods

Several other forensic feature comparison techniques are on the horizon, beyond the capacity of this reference guide to discuss. For example, “microbial forensics” is a biometric comparison technique that is increasingly used where DNA typing is not possible.⁴⁰⁹ Additionally, digital forensics—from cell phone extraction to cell site location information to hash-value authentication of documents—is a significant area of *Daubert* litigation, although digital forensics is not typically referred to as a feature comparison technique. Digital evidence is discussed in the *Reference Guide on Computer Science*, in this manual.

One technique that has been largely supplanted by mitochondrial DNA analysis, and yet which still arises in federal habeas litigation, is microscopic hair comparison. In this technique, the expert compares hair samples under a microscope and testifies whether the hair samples share a common source. Microscopic hair analysis was described by the 2009 NRC Report as a field in which there is no uniform standard on the number of features that must be congruent between hair samples for an examiner to claim a “match.” The report found that microscopic hair analysis without mtDNA testing was unreliable.⁴¹⁰ Likewise, the 2016 PCAST Report found insufficient evidence of foundational validity to support microscopic hair analysis.⁴¹¹

After the release of the 2009 NRC Report, the FBI launched a national reopening of closed cases involving hair analysis, and acknowledged misleading testimony by agents.⁴¹² Of the first 200 cases internally reviewed by the FBI, over 90% contained false or faulty testimony.⁴¹³ Most notably, Santae Tribble was imprisoned for twenty-eight years when FBI forensic analysts confused Tribble’s hair with a dog’s hair.⁴¹⁴ In partnership with the DOJ, the National Association of Criminal Defense Attorneys, and the Innocence Project, the FBI agreed to

409. See, e.g., Audrey Gouello et al., *Analysis of Microbial Communities: An Emerging Tool in Forensic Sciences*, 12 *Diagnostics* 1 (2021), <https://doi.org/10.3390/diagnostics12010001> (discussing the techniques for identifying the anatomical origin of a microbial trace, and the forensic possibilities for different types of microbial evidence).

410. See 2009 NRC Report at 161, *supra* note 37.

411. See 2016 PCAST Report, *supra* note 20.

412. See, e.g., Press Release, Innocence Project, *Innocence Project and NADCL Announce Historic Partnership with the FBI and Department of Justice on Microscopic Hair Analysis Cases* (July 18, 2013), <https://perma.cc/35VU-ALVD>.

413. Norman L. Reimer, *The Hair Microscopy Review Project: An Historic Breakthrough for Law Enforcement and a Daunting Challenge for the Defense Bar*, *Champion*, July 2013, at 16, <https://perma.cc/679W-VS4T>. See also Press Release, FBI, *FBI Testimony on Microscopic Hair Analysis Contained Errors in at Least 90% of Cases in Ongoing Review* (Apr. 20, 2015), <https://perma.cc/Y3R4-3RV8>.

414. See Simon A. Cole et al., *Microscopic Hair Comparison Analysis and Convicting the Innocent*, Nat’l Registry of Exonerations (Dec. 2023), <https://perma.cc/WD7V-EDRD>.

provide free DNA testing of any remaining evidence in the acknowledged cases. The DOJ agreed to waive any statute of limitation barriers. State initiatives followed in its wake.⁴¹⁵

Finally, a few notes about facial recognition methods are in order, given their already wide investigative use and the fact that they raise issues similar to those raised by older, relatively subjective non-DNA feature comparison methods. Facial recognition software (or facial recognition technology, FRT) compares two images to determine whether the same person is present in each image.⁴¹⁶ A probe photo, such as a still from a surveillance video, can be uploaded to a police database of photos.⁴¹⁷ The facial recognition software then provides several possible “matches,” and a human police officer uses those possible “matches” as investigative leads.⁴¹⁸

Police use of FRT and databases in investigations is now commonplace.⁴¹⁹ The FBI, for example, routinely runs face recognition searches through its agency system.⁴²⁰ Federal and state databases are also not limited to mug shots; now, nearly half of American adults are in a law enforcement agency’s facial recognition network, through the use of state driver’s license databases or other identification photos.⁴²¹

Crucially, FRT cannot yet “match” two photographs for use as identification evidence at a trial, because of its current accuracy limitations,⁴²² particularly when used to identify people of color.⁴²³ According to a 2019 NIST study evaluating the effects of race, age, and sex on facial recognition software, the

415. See *id.* at 13 n.30 (“We believe some form of state review exists in at least the following states: Arizona, Arkansas, California, Colorado, Connecticut, Florida, Illinois, Iowa, Kansas, Massachusetts, Missouri, Nebraska, New York, North Carolina, Pennsylvania, Texas, and Virginia.”). See also FBI, *Director Comey Letter to Additional Governors on State Reviews* (June 10, 2016), <https://perma.cc/ZJ3Y-3AL6>.

416. Kaitlin Jackson, *Challenging Facial Recognition Software in Criminal Court*, *Champion*, July 2019, at 14, <https://perma.cc/6HQ9-GBC6>.

417. *Id.* The database can include civilian photos from the Department of Motor Vehicles.

418. *Id.*

419. See, e.g., Hannah Bloch-Wehba, *Visible Policing: Technology, Transparency, and Democratic Control*, 109 *Calif. L. Rev.* 917, 921, 933 (2021) (noting that the extent of police use of facial recognition technology is both common and likely greater than the public is aware because of lack of transparency over its use); Claire Garvie et al., *The Perpetual Line-Up: Unregulated Police Face Recognition in America*, *Geo. L. Ctr. on Priv. & Tech.* (Oct. 18, 2016), <https://perma.cc/SC56-9UJQ>.

420. See generally Garvie, *supra* note 419.

421. *Id.*

422. See, e.g., *People v. Collins*, 15 N.Y.S.3d 564, 575 (Sup. Ct. 2015) (noting that facial recognition software, like certain cutting-edge DNA techniques, “can aid an investigation, but are not considered sufficiently reliable to be admissible at a trial”).

423. Use of facial recognition software has been shown to disproportionately affect Black Americans, Asian Americans, and Native Americans. See Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (2018); Ruha Benjamin, *Race After Technology: Abolitionist Tools for the New Jim Code* (2019).

software has been least accurate in identifying Black women, even misidentifying their gender.⁴²⁴ Two recent high-profile cases involved false accusations of crime against people of color.⁴²⁵ In response, software datasets are growing to reflect diversity, in order to diminish error rates.⁴²⁶ Still, while making datasets more representative addresses one source of error, another source of error can be the comparison of features by human examiners. Most recently, the National Academy of Sciences issued a comprehensive report on FRT, including a discussion of its uses, limitations, and recommendations for best practices.⁴²⁷

FRT output as of this writing has not yet been admitted as identification evidence at trial, and at least one trial court has observed that there is “no agreement in a relevant community of technological experts that [FRT] matches are sufficiently reliable to be used in court as identification evidence.”⁴²⁸ Nonetheless, courts should be aware of its investigative use and keep abreast of future developments in the field. As the 2024 NAS Report notes, judges should be ready not only to determine whether FRT as an identification technique is foundationally reliable, but also whether it is “applied reliably by an appropriately trained, competent analyst.”⁴²⁹ Likewise, speaker recognition technology, as of this writing, is being used to investigate crimes such as false marine distress calls, but is rarely as yet used in criminal cases as evidence.⁴³⁰

424. NIST, *NIST Study Evaluates Effects of Race, Age, Sex on Face Recognition Software* (Dec. 19, 2019), <https://perma.cc/C5KV-HYLS>.

425. See Kashmir Hill, *Eight Months Pregnant and Arrested After False Facial Recognition Match*, N.Y. Times, Aug. 6, 2023, <https://www.nytimes.com/2023/08/06/business/facial-recognition-false-arrest.html> (Porcha Woodruff falsely accused of robbery and carjacking); Tate Ryan-Mosley, *The New Lawsuit That Shows Facial Recognition Is Officially a Civil Rights Issue*, MIT Tech. Rev., Apr. 14, 2021, <https://perma.cc/89WV-E2LL> (noting false theft and burglary accusation against Robert Williams in Detroit).

426. Additionally, there is a NIST OSAC Facial Identification Subcommittee that works “to develop consensus standards and guidelines for the image-based comparisons of human facial features and to provide recommendations for the research and development necessary to advance the state of the science.” See NIST, *Facial Identification Subcommittee*, <https://perma.cc/XJ4K-SPLQ>.

427. Nat’l Acad. of Scis., Eng’g & Med. (NAS), *Facial Recognition Technology: Current Capabilities, Future Prospects, and Governance* (2024), <https://doi.org/10.17226/27397> [hereinafter 2024 NAS Report].

428. *People v. Reyes*, 133 N.Y.S.3d 433, 436–37 (Sup. Ct. 2020).

429. 2024 NAS Report, *supra* note 427, at 103.

430. See generally Peter Andrey Smith, *Can We Identify a Person from Their Voice?*, IEEE Spectrum, Apr. 15, 2023, <https://perma.cc/4KUL-Q8VM> (noting current investigative uses of voice recognition technology but also that it is not yet used in criminal trials in the United States).

Special Issues with Machine-Generated Feature Comparison Evidence

Feature comparison methods are being increasingly automated. Previous sections have already described some of these turns toward, or research to eventually enable, expert systems in place of human analysis in feature comparison. This section briefly addresses unique conceptual and legal issues raised by this automation.

First, a brief note about definitions. By machine-generated feature comparison evidence, we mean the conveyance by a machine of a conclusion about whether or how two features or patterns “match,” such as a software program’s reported “likelihood ratio” comparing the chance of seeing the evidence sample under competing hypotheses about whether the defendant is the source of the evidence. In that case, the expert system itself is generating and reporting the information output, just like a human expert might make an assertion in a report or on the witness stand. Machine-generated feature comparison is therefore distinct from mere electronic storage of, or conduits for, information. Courts have admitted machine-stored information since at least the 1970s.⁴³¹ Indeed, electronically stored information (ESI) has now been officially incorporated into the rules of civil discovery.⁴³² Likewise, courts routinely admit emails and social media posts made by people and communicated via computers. To be sure, such evidence raises potential authentication questions (such as whether a purported email is actually written by a particular declarant).⁴³³ But ESI generally does not raise novel issues related to the reliability of the information itself, such as whether an assertion in an email is true.

Types of Machine-Generated Feature Comparison Evidence

The past decade has seen a precipitous rise in feature comparison techniques automated or at least facilitated by computer software. Techniques that were previously the province only of human examiners, such as firearm and toolmark

431. See, e.g., *United States v. Liebert*, 519 F.2d 542, 543 (3d Cir. 1975) (admitting list of persons not filing tax returns, stored in IRS computers).

432. See Fed. R. Civ. P. 34.

433. See Fed. R. Evid. 902(13), (14) & advisory committee’s notes to 2017 amendments. For an example of a successful authentication challenge to ESI, see *United States v. Vayner*, 769 F.3d 125 (2d Cir. 2014) (holding that admission of social media page and contents against defendant was reversible error because not properly authenticated as having been posted by him).

comparison,⁴³⁴ handwriting comparison,⁴³⁵ shoemark comparison,⁴³⁶ latent print comparison,⁴³⁷ facial recognition,⁴³⁸ and arson analysis based on burn or debris patterns,⁴³⁹ can now potentially be done by algorithm. The Government Accounting Office issued a report in 2021 on issues raised by automated forensic methods—in particular DNA, latent print analysis, and facial recognition.⁴⁴⁰ This reference guide focuses on non-DNA comparison disciplines. Nonetheless, because existing cases addressing forensics algorithms are all in the DNA context, judges should have a general understanding of such cases. Competing software programs now purport to interpret complex DNA mixtures, determine whether a defendant's profile is consistent with the mixture, and report associated match statistics. Courts have generally admitted the results of such programs—typically, but not exclusively, offered by the prosecution⁴⁴¹—over *Frye/Daubert* objections.⁴⁴² The few exceptions have been cases in which a local laboratory did not perform internal validation studies before using the software,⁴⁴³ where a mixture had a particularly

434. See generally, Eric Hare, Heike Hofmann & Alicia Carriquiry, *Automatic Matching of Bullet Land Impressions*, 11 *Annals Applied Stat.* 2332 (2017), <https://doi.org/10.1214/17-AOAS1080> (creating an open-source algorithm for automatically comparing certain striations, after removing class characteristics, on fired ammunition).

435. See, e.g., Amy M. Crawford, Danica M. Ommen & Alicia L. Carriquiry, *A Statistical Approach to Aid Examiners in the Forensic Analysis of Handwriting*, 68 *Annals Applied Stat.* 1768 (2023), <https://doi.org/10.1111/1556-4029.15337>.

436. Soyoun Park & Alicia Carriquiry, *An Algorithm To Compare Two-Dimensional Footwear Outsole Images Using Maximum Cliques and Speeded-Up Robust Feature*, 13 *Stat. Analysis & Data Mining: ASA Data Sci. J.* 188 (2020), <https://doi.org/10.1002/sam.11449>.

437. See Simon A. Cole et al., *Beyond the Individuality of Fingerprints: A Measure of Simulated Computer Latent Print Source Attribution Accuracy*, 7 *L. Prob. & Risk* 165, 166 (2008), <https://doi.org/10.1093/lpr/mgn004> (explaining how the AFIS database returns several “candidate” prints to the analyst).

438. See, e.g., *Lynch v. State*, 260 So.3d 1166 (Fla. Dist. Ct. App. 2018).

439. See, e.g., Sander Korver et al., *Artificial Intelligence and Thermodynamics Help Solving Arson Cases*, 10 *Sci. Reps.* 20502 (2020), <https://www.nature.com/articles/s41598-020-77516-x>.

440. See Karen L. Howard, *Forensic Technology: Algorithms Offer Benefits for Criminal Investigations, but a Range of Factors Can Affect Outcomes*, Gov't Accountability Off. (Jan. 24, 2024), <https://perma.cc/SW4Q-TX3Z>.

441. One such program, TrueAllele, has been used by a handful of defendants to demonstrate their innocence. See generally, *TrueAllele Helps Free Innocent Indiana Man After 24 Years in Prison*, Cybergenetics, Apr. 26, 2016, <https://perma.cc/RA7Q-XV92>.

442. See John S. Hausman, *Lost Shoe Led to Landmark DNA Ruling—And Now, Nation's 1st Guilty Verdict*, MLive, <https://perma.cc/UA3W-F8XF> (reporting a conviction in Michigan as the first in the United States to be based in part on the STRmix software after the defense contested its admissibility); *Trials*, Cybergenetics, <https://perma.cc/V9ZK-MFP6> (listing over fifty cases in which TrueAllele has been admitted).

443. See Decision & Order on DNA Analysis Admissibility, *People v. Hillary*, Indictment No. 2015-15 (N.Y. Cnty. Ct. Aug. 26, 2016), <http://perma.cc/3TFA-8VA5> (excluding STRmix results under *Frye*).

high number of contributors,⁴⁴⁴ where the evidence involved a minor contributor to a mixture,⁴⁴⁵ or where the software was discontinued.⁴⁴⁶ In October 2019, in *United States v. Gissantaner*, a federal district judge ruled the results of one program inadmissible in a case involving “low copy number” DNA,⁴⁴⁷ but the Sixth Circuit reversed the following spring.⁴⁴⁸

One particular 2016 case, although involving DNA, is important to be aware of, as it will likely be invoked in future challenges to results of other automated forensic analyses. In a high-profile New York homicide case, the two main DNA programs—STRmix™ and TrueAllele®—reached very different conclusions. The case involved the strangulation of a young boy. Police suspicion fell upon Oral Hillary, a former college soccer coach who had dated the boy’s mother.⁴⁴⁹ Another former boyfriend, a deputy sheriff who had been physically violent with the mother, was cleared of suspicion. No physical evidence linked either man to the scene. But analysts eventually collected a DNA mixture, too complex to be analyzed by humans, from under the boy’s fingernail. Police in 2013 gave the DNA fingernail data to the proprietor of TrueAllele. In 2014, TrueAllele reported “no statistical support for a match” with Hillary.⁴⁵⁰ Indeed the TrueAllele CEO would later insist the evidence suggested Hillary was excluded as a contributor, based on the reported likelihood ratio.⁴⁵¹ A year later, a new district attorney had the DNA data analyzed through STRmix, which reported that Hillary was 300,000 times more likely than a random person to have contributed to the mixture.⁴⁵² A trial judge excluded the STRmix results because of a lack of internal validation,⁴⁵³ and

444. See Order Excluding DNA Evidence, *United States v. Alfonzo Williams*, Case No. 3:13-cr-00764-WHO-1 (N.D. Cal. Apr. 29, 2019) (Orrick, J.) (excluding results of “Bullet-proof” genotyping software in case with potentially more than four contributors).

445. See *United States v. Lewis*, 442 F. Supp. 3d 1122 (D. Minn. 2020).

446. See Shayna Jacobs, *Judge Tosses Out Two Types of DNA Evidence Used Regularly in Criminal Cases*, N.Y. Daily News, Jan. 5, 2015, <https://perma.cc/A9ZL-BMJ6> (reporting a Brooklyn judge excluded results from low copy number DNA testing and Forensic Statistical Tool testing).

447. *United States v. Gissantaner*, 417 F. Supp. 3d 857 (W.D. Mich. 2019) (Neff, J.), *rev’d*, 990 F.3d 457 (6th Cir. 2021).

448. *Gissantaner*, 990 F.3d 457. Rehearing en banc was denied, *id.*, and the case appears to have since ended in a change of plea. See <https://perma.cc/E765-9FW6>.

449. See, e.g., Jesse McKinley, *Tensions Simmer over Race as Town Reels from Boy’s Killing*, N.Y. Times, Mar. 5, 2016, at A1, <https://perma.cc/AE3L-NPVW>.

450. W.T. Eckert, *Hillary Trial Slated for Aug. 1*, Watertown Daily Times (Mar. 2, 2016).

451. See John Buckleton, Memorandum, *People v. Hillary* (2017), <https://perma.cc/H3UE-GT7U> (proprietor of STRmix explaining the discrepancies between STRmix and TrueAllele likelihood ratios in the *Hillary* case).

452. Notice of Motion to Preclude at 10, *People v. Hillary*, Indictment No. 2015-15 (N.Y. Cnty. Ct. May 31, 2016), <https://perma.cc/2H69-DJJT>.

453. See Decision & Order, *People v. Hillary*, Indictment No. 2015-15 (N.Y. Cnty. Ct. Aug. 26, 2016) (Catena, J.).

Hillary was acquitted.⁴⁵⁴ The creator of STRmix has since published a memorandum online, explaining how STRmix’s approach to the evidence in *Hillary* is different from TrueAllele’s approach.⁴⁵⁵ The case is an example of two programs, both deemed reliable by courts under *Daubert* or *Frye*, coming to significantly different conclusions based on the same forensic data.⁴⁵⁶ Judges should also be aware of a 2021 NIST study exploring the reliability of probabilistic genotyping software, given that its findings and recommendations might be relevant to evaluation of other automated methods.⁴⁵⁷

Recurring Legal Issues

This section offers an overview of legal issues potentially raised by machine-generated feature comparison evidence,⁴⁵⁸ some of which are already the subject of federal litigation.

Rule 702/Reliability Issues

If a computer program or machine output is the method upon which a human feature comparison expert’s opinion is based, then the algorithm might be subject to *Daubert* scrutiny.⁴⁵⁹

454. Jesse McKinley, *Race, Jilted Love and Acquittal in Boy’s Killing*, N.Y. Times, N.Y. Ed., at A1 (Sept. 29, 2016).

455. See Buckleton, *supra* note 451.

456. The study emphasized the importance, in determining software accuracy in a given case, of validation studies covering the particular “factor space” that a case falls into (such as, in the DNA mixture context, the quantity of DNA, number of contributors, etc.). See John M. Butler et al., NIST, DNA Mixture Interpretation: A NIST Scientific Foundation Review (2021), <https://doi.org/10.6028/NIST.IR.8351-draft> (noting that report is in draft form but that comment period has closed).

457. *Id.*

458. For a general discussion of legal issues related to algorithmic feature comparison evidence, see generally Andrea Roth, *The Use of Algorithms in Criminal Adjudication*, in Cambridge Handbook on the Law of Algorithms 407 (Woodrow Barfield ed., 2020), <https://doi.org/10.1017/9781108680844> (describing the rise of algorithmic proof in criminal cases, and the legal and ethical issues raised); Edward K. Cheng & G. Alexander Nunn, *Beyond the Witness: Bringing a Process Perspective to Modern Evidence Law*, 97 Tex. L. Rev. 1077 (2019) (discussing how evidence law and Confrontation Clause jurisprudence should accommodate computer-generated evidence); Andrea Roth, *Machine Testimony*, 126 Yale L.J. 1972 (2017) (describing the rise of machine-generated conveyances of information as proof, and testimonial safeguards that might be deployed to better regulate such proof in a manner analogous to human assertions).

459. See Fed. R. Evid. 702 (applying only to expert “witnesses” who testify). *Cf.* *People v. Lopez*, 286 P.3d 469, 478 (Cal. 2012) (portion of spectrometer readings offered to prove blood-alcohol concentration, without additional sworn testimony or certification by human expert, not

Reliability concerns have arisen with respect to several types of machine-generated conclusions both within and outside the feature comparison context, including driving estimates,⁴⁶⁰ blood alcohol testing,⁴⁶¹ translation of the federal sentencing guidelines into computer code,⁴⁶² Apple’s “Find My iPhone” tracking when used in theft and robbery cases,⁴⁶³ “minor miscode[s]” in DNA software,⁴⁶⁴ and inaccurate allelic frequencies used by software to generate DNA match statistics.⁴⁶⁵ Additionally, as discussed in the section titled “Facial Recognition Software and Other Emerging Methods” above, machine learning algorithms can misidentify or misclassify people or events because of limitations in the training dataset or other analytical problems.⁴⁶⁶ Machines learn how to characterize new data by training on data already categorized by a person or the machine itself. If the dataset is too small or too unrepresentative of future items to be classified, the algorithm might infer a pattern or linkage in the training data that does not actually mirror real life (overfitting) or might try to account for too many variables, rendering the training data inadequate for learning (the so-called curse of dimensionality).⁴⁶⁷

With respect to issues that might arise in a *Daubert* hearing related to software results, some commentators have argued that validation studies alone might be insufficient to scrutinize certain algorithmic proof. For example, while validation studies might show that a certain interpretive software program boasts a

subject to the Confrontation Clause because not the statement of a human expert, over dissent by Justice Liu).

460. See, e.g., *Gianniney v. State*, 962 A.2d 229, 232 (Del. 2008) (excluding Mapquest driving estimates as inadmissible hearsay and noting reliability concerns).

461. See Andrea Roth, *Trial by Machine*, 104 Geo. L.J. 1245, 1271–72 (2016) (citing sources with respect to problems with Intoxilyzer 8000 and 5000 readings, due to programming errors); Roth, *Machine Testimony*, *supra* note 458, at 1995 (discussing litigation over code errors with the Alcotest 7110).

462. See Steven R. Lindemann, Commentary, *Published Resources on Federal Sentencing*, 3 Fed. Sent’g Rep. 45, 45–46 (1990), <https://doi.org/10.2307/20639272>.

463. See Lawrence Mower, *If You Lose Your Cellphone, Don’t Blame Wayne Dobson*, Las Vegas Rev.-J., Jan. 13, 2013, <https://perma.cc/2G7G-89QX>.

464. Roth, *Trial by Machine*, *supra* note 461, at 1276.

465. See Notice of Amendment of the FBI’s STR Population Data Published in 1999 and 2001, FBI (2015), <https://perma.cc/TDF7-WMLA>.

466. See also *Face Recognition*, Elec. Frontier Found., <https://perma.cc/W648-PQWF> (citing Brendan F. Klare et al., *Face Recognition Performance: Role of Demographic Information*, 7 IEEE Transactions Info. Forensics & Sec. 1789 (2012), <https://doi.org/10.1109/TIFS.2012.2214212> (showing higher false positive rates for African Americans)).

467. See generally *The Discipline of Organizing: Informatics Edition* (Robert J. Glushko ed., 4th ed. 2013); Harry Surden, *Machine Learning and Law*, 89 Wash. L. Rev. 87 (2014); Pedro Domingos, *The Master Algorithm: How the Quest for the Ultimate Learning Machines Will Remake Our World* (2015); Alice Zheng, *Evaluating Machine Learning Models: A Beginner’s Guide to Key Concepts and Pitfalls* (2015).

low false positive rate, such studies cannot so easily determine whether a reported likelihood ratio is inaccurate:

Laboratory procedures to measure a physical quantity such as a concentration can be validated by showing that the measured concentration consistently lies with an acceptable range of error relative to the true concentration. Such validation is infeasible for software aimed at computing a[] [likelihood ratio] because it has no underlying true value (no equivalent to a true concentration exists). The [likelihood ratio] expresses our uncertainty about an unknown event and depends on modeling assumptions that cannot be precisely verified in the context of noisy [crime scene profile] data.⁴⁶⁸

In sum, judges faced with determining foundational validity and validity as applied of machine-generated feature comparison conclusions will need to consider separately whether the software's classifications are reliable (e.g., identification versus nonidentification; consistent versus inconsistent) and whether the software's reported "scores" are reliable (such as a match statistic or score where the ground truth is not known). Judges will face the same questions as with other feature comparison methods in terms of what counts as a well-designed validation study, and how many well-designed studies are necessary, to determine a program's error rate. They will also need to decide whether some issues (such as the reliability of reported scores that cannot be tested through traditional black box validation studies) require greater access to the source code underlying the software, a discovery issue discussed below.

Hearsay and Confrontation

Some litigants have tried to argue, most unsuccessfully, that a machine-generated conclusion offered as evidence is "hearsay" and therefore inadmissible unless it comes within an exception to the rule against hearsay. Hearsay, under Federal Rules of Evidence 801 and 802, is an out-of-court "assertion" by a "person," "intended" as an assertion, and offered to prove the truth of the matter asserted.⁴⁶⁹ By the explicit language of these rules, as numerous courts have now held,⁴⁷⁰ a factual claim rendered by a machine rather than a person cannot be

468. Christopher D. Steele & David J. Balding, *Statistical Evaluation of Forensic DNA Profile Evidence*, 1 Ann. Rev. Stat. & Its Application 361, 380 (2014), <https://doi.org/10.1146/annurev-statistics-022513-115602>.

469. See, e.g., Fed. R. Evid. 801, 802 (defining hearsay as an out-of-court "statement," "intended" as an assertion, offered to prove the truth of the matter asserted, and stating that hearsay is presumptively inadmissible).

470. See, e.g., *People v. Lopez*, 286 P.3d 469, 478 (Cal. 2012) (holding that gas chromatograph "raw" data is not hearsay); *United States v. Lizarraga-Tirado*, 789 F.3d 1107, 1109 (9th Cir. 2015) (holding that Google Earth output is not hearsay).

hearsay. To be sure, some courts used to treat computer-generated results as hearsay and admit them under the “business records” hearsay exception for records created in the regular course of business.⁴⁷¹ But commentators were critical of this practice, pointing out that these courts were conflating electronically stored records, inputted by humans but stored by computers, with electronically generated records, where the computer itself creates and reports information.⁴⁷²

Some courts and commentators have alternatively suggested that an algorithm’s conclusions might be the hearsay statements of the algorithm’s creator, such as a computer programmer, and that machine-generated proof is admissible so long as the programmer is subject to cross-examination.⁴⁷³ Others have argued against this view of algorithmic evidence, noting that the programmer has not uttered the resulting information, and might not even fully understand the thousands of steps taken by the program to generate the information, particularly in more advanced forms of artificial intelligence (AI).⁴⁷⁴

Some litigants and commentators have further argued that admission of machine-generated conclusions without sufficient opportunity for adversarial scrutiny may violate the Sixth Amendment’s Confrontation Clause,⁴⁷⁵ which guarantees an accused the right “to be confronted with the witnesses against him.” To be sure, under existing U.S. Supreme Court precedent, the right of confrontation applies only to “testimonial hearsay,”⁴⁷⁶ which presumably only covers certain human assertions. While the Court has not squarely addressed the question, Justice Sotomayor has suggested in a concurring opinion that “raw data” from a machine might not be “testimonial.”⁴⁷⁷ Not surprisingly, then,

471. Apparently, this approach was buttressed by an oft-cited 1974 article arguing that “[c]omputer generated evidence will inevitably be hearsay.” Jerome J. Roberts, *A Practitioner’s Primer on Computer-Generated Evidence*, 41 U. Chi. L. Rev. 254, 272 (1974).

472. See generally Adam Wolfson, Note, “Electronic Fingerprints”: *Doing Away with the Conception of Computer-Generated Records as Hearsay*, 104 Mich. L. Rev. 151 (2005) (noting, and criticizing, courts’ treatment of computer-generated business records as hearsay).

473. See, e.g., *People v. Wakefield*, 38 N.Y.3d 367, 385–86 (2022) (holding that admission of TrueAllele results did not violate the Confrontation Clause so long as the creator, Mark Perlin, was on the witness stand); *United States v. Washington*, 498 F.3d 225, 229 (4th Cir. 2007) (explaining defendant’s argument that the “raw data” of a chromatograph was the “hearsay” of the “technicians” who tested the defendant’s blood sample for PCP and alcohol using the machine); Karen Neville, *Programmers and Forensic Analyses: Accusers Under the Confrontation Clause*, 10 Duke L. & Tech. Rev. 1, 8–9 (2011) (arguing that “the programmer” is “the ‘true accuser’—not the machine merely following the protocols he created”).

474. See, e.g., Roth, *Machine Testimony*, *supra* note 458, at 1986. See generally Cheng & Nunn, *supra* note 458 (describing potential inaccuracies of “process-based” proof and suggesting enhanced pretrial discovery, not cross-examination of an individual expert, as the most appropriate safeguard).

475. See, e.g., *Wakefield*, 38 N.Y.3d at 385 (arguing that admission of TrueAllele results without disclosure of the source code violated the Confrontation Clause).

476. *Crawford v. Washington*, 541 U.S. 36 (2004).

477. *Bullcoming v. New Mexico*, 564 U.S. 647, 674 (2011) (Sotomayor, J., concurring).

most courts have thus far rejected arguments that the admission of a machine's claim without disclosing certain information about a machine's processes (such as source code) violates the Confrontation Clause.⁴⁷⁸ Still, several commentators have argued that any approach categorically exempting machine assertions from the right of confrontation may soon become untenable, especially as machine-generated conclusions become ever more complex, opaque, and interpretive, akin to human expert testimony.⁴⁷⁹

Discovery Issues

A final recurring legal issue related to machine-generated feature comparison evidence is disputes over pretrial requests for disclosure of the source code or for access to a research license to conduct software audits. The dispute stems from the fact that many (though not all) of the programs underlying machine-generated proof are proprietary,⁴⁸⁰ and program developers often argue, in

478. See, e.g., *Wakefield*, 38 N.Y.3d at 385 (2022) (holding that admission of TrueAllele results without disclosure of the source code did not violate the Confrontation Clause); *People v. Lopez*, 286 P.3d 469, 478 (Cal. 2012) (holding that gas chromatograph "raw" data is not hearsay and thus its admission did not implicate the Confrontation Clause). *But see* *People v. Chubbs*, No. B258569, 2015 WL 139069, at *4 (Cal. Ct. App. Jan. 9, 2015) (noting that the trial court had invoked the Confrontation Clause in ordering disclosure of source code to facilitate cross-examination of programmer); *Order on Procedural History and Case Status in Advance of May 25, 2016 Hearing*, *United States v. Michaud*, No. 3:15-CR-05351RJB, 2016 WL 337263 (W.D. Wash. Jan. 28, 2016) (noting due process right to examine source code of government's Network Investigative Technique (NIT) used to hack defendant's computer).

479. See Andrea Roth, *What Machines Can Teach Us About "Confrontation,"* 60 *Duquesne L. Rev.* 210 (2022) (explaining that a view of confrontation as merely guaranteeing cross-examination at trial is ahistorical); Cheng & Nunn, *supra* note 458 (arguing that machine-generated proof cannot be exempt from confrontation); Roth, *Machine Testimony*, *supra* note 458 (same); Erin Murphy, *The Mismatch Between Twenty-First-Century Forensic Evidence and Our Antiquated Criminal Justice System*, 87 *S. Calif. L. Rev.* 633, 657–58 (2014); David Alan Sklansky, *Hearsay's Last Hurrah*, 2009 *Sup. Ct. Rev.* 1, 67 (2009) (urging a view of confrontation as "a meaningful opportunity to test and to challenge the prosecution's evidence"); see also *id.* at 7 (quoting Daniel H. Pollitt, *The Right of Confrontation: Its History and Modern Dress*, 8 *J. Pub. L.* 381, 402 (1959)) ("The Confrontation Clause could be read broadly to guarantee criminal defendants a meaningful opportunity to challenge—to know, to examine, to explain, and to rebut—the proof offered against them.").

480. See, e.g., *State v. Loomis*, 881 N.W.2d 749 (Wis. 2016) (denying litigants access to source code of actuarial tool used in parole hearing); *Chubbs*, 2015 WL 139069, at *8–9 (noting that DNA mixture interpretation software TrueAllele is proprietary); *Order, United States v. Michaud*, No. 3:15-CR-05351RJB (W. Dist. Wash. May 18, 2016) (noting that government's malware used in child pornography investigation was proprietary); Luke Broadwater & Scott Calvert, *City in \$2 Million Dispute With Xerox Over Camera Tickets*, *Balt. Sun*, Apr. 24, 2013, <https://perma.cc/T2X7-F2PS> (noting that Xerox refused to disclose source code for red light camera system). *But see* Hare, Hofmann & Carriquiry, *supra* note 434 (creating an open-source algorithm for automatically comparing certain striations, after removing class characteristics, on fired ammunition, and noting the benefits of open-source software for transparency and improvement).

response to discovery requests for information like source code, that such information is protected by a “trade secrets privilege.”⁴⁸¹ Some have suggested that a trade secret privilege might be necessary to incentivize development of high-quality algorithms for use in criminal justice,⁴⁸² while others have argued that such a privilege is inappropriate and unnecessary.⁴⁸³ In particular, Professor Rebecca Wexler has argued that the emergence of a trade secrets privilege in criminal cases was an historical accident, and that substantive trade secret doctrine offers software developers sufficient protection to incentivize innovation. Until recently, courts appeared to have largely accepted proprietors’ claims of privilege.⁴⁸⁴ At least two courts have granted defense requests for source code in cases involving DNA software, however.⁴⁸⁵ Meanwhile, U.S. Representative Mark Takano (D-CA) introduced a bill (originally introduced in 2019) titled “Justice in Forensic Algorithms Act of 2024” that would require disclosure of source code and remove any trade secret privilege over such code.⁴⁸⁶

Admission of machine-generated feature comparison conclusions might raise other discovery issues as well. For example, a party might ask for pretrial access to a program to manipulate its inputs, in the same way that a party might seek to depose or interview an opposing party’s expert witness before trial.⁴⁸⁷ Next, parties might ask for access to the prior statements or runs of software programs in a given case. While there is no Jencks Act for machine conveyances, judges will have to decide whether to grant access to such statements as fairness demands.⁴⁸⁸ In addition, judges might face requests by a party for access to the

481. See, e.g., Roth, *Machine Testimony*, *supra* note 458, at 2028 (discussing trade secrets); Christian Chessman, Note, *A “Source” of Error: Computer Code, Criminal Defendants, and the Constitution*, 105 Calif. L. Rev. 179, 212 (2017) (discussing the trade secret privilege in criminal cases with respect to source code).

482. See, e.g., Edward J. Imwinkelried, *Computer Source Code: A Source of the Growing Controversy over the Reliability of Automated Forensic Techniques*, 66 DePaul L. Rev. 97 (2016).

483. See Chessman, *supra* note 481, at 212; Rebecca Wexler, *Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System*, 70 Stan. L. Rev. 1343 (2018); Jason Tashea, *Trade Secret Privilege Is Bad for Criminal Justice*, A.B.A. J., July 30, 2019, <https://perma.cc/CKY9-HLVN>.

484. Wexler, *supra* note 483, at 1352–53.

485. The New Jersey case is *State v. Pickett*, 246 A.3d 279 (N.J. Super. Ct. App. Div. 2021). The state has since withdrawn its intent to introduce the TrueAllele results, and thus this litigation appears to be moot. The other case is *United States v. Ellis*, No. 19–369, 2021 WL 1600711 (W.D. Pa. Apr. 23, 2021). As of this writing, the parties had agreed upon terms of a protective order and review by experts had just begun.

486. See H.R. 7394, 118th Cong. (Feb. 15, 2024), <https://www.congress.gov/bill/118th-congress/house-bill/7394/committees>.

487. Cf. Jennifer L. Mnookin, *Repeat Play Evidence: Jack Weinstein, “Pedagogical Devices,” Technology, and Evidence*, 64 DePaul L. Rev. 571, 588 (2015) (arguing that demonstrative evidence in the form of complex algorithms should have this as a condition of admissibility).

488. See, e.g., Roth, *Machine Testimony*, *supra* note 458, at nn.411–19 & accompanying text (explaining why the *Jencks* case (*Jencks v. United States*, 353 U.S. 657 (1957)) can and should apply to machines and why it would be consistent with the *Jencks* Act).

Reference Guide on Forensic Feature Comparison Evidence

data on which a machine-learning algorithm was “trained” to classify future objects or events. A party might seek this dataset to argue that the algorithm was trained on data that was not representative of the person, object, or event being classified in the case (such as facial recognition software claiming to have identified an African American as having been at a crime scene, where the software was trained largely on white faces). The proponent of the evidence might argue that the data are covered by privacy laws and cannot be disclosed. Just as judges have always had to grapple with access to privacy-protected data in contexts like a victim’s medical records, judges will need to consult applicable data-protection laws, with the accused’s constitutional right to present a defense in mind, in resolving these issues.⁴⁸⁹

489. For a brief discussion of intellectual property and data privacy issues related to machine learning training data, see Brittany Bacon et al., *Training a Machine Learning Model Using Customer Proprietary Data: Navigating Key IP and Data Protection Considerations*, in Pratt’s Privacy and Cybersecurity Law Report, LexisNexis 233 (Steven A. Meyerowitz et al. eds., Oct. 2020).

Glossary of Terms

This glossary has been adapted from a general literature review related to the topics included in this reference guide. An ongoing project by NIST's Organization of Scientific Area Committees (OSAC) is also creating a list of preferred terms for recurring concepts in forensic science.⁴⁹⁰

ACE-V. Acronym for the method many feature comparison analysts use and refer to as Analysis, Comparison, Evaluation, and Verification.

AFIS. Acronym for Automated Fingerprint Identification System used by law enforcement to rapidly screen large numbers of prints in a national database, which are then evaluated by analysts. While AFIS is useful for screening, the system is not currently able to determine that a particular pair of prints has a common source.

algorithm. A list of steps capable of being performed by a machine.

“black box” study. A validation study designed to determine the extent to which a forensic examiner applying a feature comparison method reaches the correct conclusion (where the ground truth is known), without having to know anything about how the method itself—which may well remain a “black box”—works.

class characteristics. Characteristics believed to be shared by a group of persons or objects (e.g., ABO blood types).

contextual bias. Bias that occurs when a forensic examiner's judgment while engaged in feature comparison is influenced by irrelevant information about the facts of a case.

double-“blind” or double-anonymized research or testing. Research or testing in which the respondent and the interviewer are not given information that will alert them to the anticipated or preferred pattern of response. While the phrases “blind” and “double-blind” have been eliminated by many users over concerns that such language is ableist, judges will still frequently encounter these terms in literature discussing feature comparison disciplines.

false negative rate. The rate at which an examiner, using a feature comparison method, erroneously concludes that two features or patterns are inconsistent when, in fact, they are consistent. Note that there is a debate in the literature regarding how to incorporate an examiner's determination of “inconclusive” (in a testing context, not in casework), when the ground truth is that two features or patterns are consistent, into a false negative rate.

490. See NIST, *OSAC Lexicon*, <https://perma.cc/VKY4-KEND>.

false positive rate. The rate at which an examiner, using a feature comparison method, erroneously concludes that two features or patterns are consistent when, in fact, they are not consistent. Note that there is a debate in the literature regarding how to incorporate an examiner's determination of "inconclusive" (in a testing context, not in casework), when the ground truth is that two features or patterns are inconsistent, into a false positive rate.

FDE. Acronym for Forensic Document Examiners, who compare known and unknown writing samples and perform various tests on documents, such as whether a particular ink formulation existed on the purported date of a writing. Also referred to as Questioned Document Examiners (QDE).

fingerprint conclusions. A fingerprint examiner's conclusion as to the strength of the evidence in supporting an inference about whether two impressions came from the same source. Some laboratories, such as the Department of Justice, guide examiners to use one of three conclusions: source identification, a source exclusion, or inconclusive determination.

fingerprint inconclusive determination. The label some examiners use to describe their determination that there is "insufficient quantity and/or clarity" between the two impressions for the examiner to arrive at a source identification or exclusion of fingerprints.

fingerprint source exclusion. The label some examiners use to describe their determination that the two sets of fingerprints did not come from the same source.

fingerprint source identification. A label some examiners use to describe their determination that two sets of fingerprint prints—known and unknown—came from the same person. Other examiners might eschew this "identification" language and instead make statements about the strength of the evidence, such as that there is "extremely strong support" that the impressions came from the same source and "extremely weak support" that they came from different sources.

individual characteristic. A characteristic believed to be unique to an object or person.

individualization. The claim that a feature comparison method can scientifically determine that a feature or pattern (such as a fingerprint, or toolmark, or bite mark) was produced by one source, to the exclusion of all other sources. This claim is notably different from the concept of "uniqueness," a claim that a particular biological trait (such as the ridges on one's fingers) is unique to each person.

latent prints. Fingerprints found at a crime scene or in another location, which may vary in quality or number. Latent prints are compared to record prints to determine whether there is consistency between the two.

likelihood ratio (LR). A ratio in which the numerator is the chance of seeing a particular type of evidence (such as a DNA “match”) given one hypothesis (e.g., that the defendant is the source of the DNA) and the denominator is the chance of seeing the evidence (e.g., the DNA “match”) given the alternative hypothesis (such as that the defendant is *not* the source of the DNA). An LR is part of Bayes’ Theorem, a means of updating a prior guess about the probability of an event with new evidence (incorporated into the LR) to arrive at a new guess (“posterior probability”). Computer programs purporting to interpret complex DNA mixtures typically report their conclusions in the form of LRs.

machine learning. An algorithm in which the computer is “trained” on a dataset (such as emails already classified as “spam” or “not spam”) such that the computer “learns” to classify objects in the future (such as determining whether future emails are “spam” versus “not spam”). More broadly, machine learning refers to any system in which a machine learns from experience and improves its performance on an assigned task over time.

OSAC. The Organization of Scientific Area Committees, an organization under the direction of the National Institute of Standards and Technology (NIST), under the United States Department of Commerce. OSAC focuses on developing and approving standards to govern forensic science service providers.

PCAST. President’s Council of Advisors on Science and Technology.

QDE. See FDE.

record prints. Fingerprints that are rolled onto a fingerprint card or digitized and scanned into a file. These prints are the ones stored in databases and against which unknown prints are compared.

reliability. The extent to which the same results are obtained in each instance in which a test or method is performed—its consistency. Reliability can be further divided into repeatability, the ability of one examiner to repeat the same test results using the same method under the same conditions; and reproducibility, the ability of another examiner to repeat the same test results using the same method under the same conditions.

source code. The code written by a computer programmer, in human-readable form, to a computer containing the instructions for what to do.

SWGDOC. The Scientific Working Group for Forensic Document Examination.

SWGFAST. The Scientific Working Group on Friction Ridge Analysis, Study, and Technology.

toolmark conclusions. A toolmark examiner’s conclusion as to the strength of the evidence in supporting an inference about whether two toolmarks came from the same source. Some laboratories, such as the Department of Justice,

Reference Guide on Forensic Feature Comparison Evidence

guide examiners to use one of three conclusions: source identification, source exclusion, or inconclusive determination. For a description of each, see Fingerprint Conclusions.

ULTR. The Department of Justice’s guidelines on Uniform Language for Testimony and Reports.

validity. The ability of a test to measure what it is supposed to measure—its accuracy. Validity includes reliability, but the converse is not necessarily true.

Reference Guide on Human DNA Identification Evidence

DAVID H. KAYE

David H. Kaye, M.A., J.D., is Regents Professor Emeritus, Arizona State University Sandra Day O'Connor College of Law and School of Life Sciences, and Distinguished Professor of Law and Academy Professor Emeritus, Pennsylvania State University School of Law.

Author's Note: Research for this reference guide was completed in 2023. I am grateful not only to the National Academies of Sciences, Engineering, and Medicine and Federal Judicial Center reviewers and staff, but also to Bruce Budowle, John Butler, Michael Coble, David Housman, Jarrah Kennedy, Swathi Kumar, and Bruce Weir for their comments on portions of a draft of this reference guide.

CONTENTS

Introduction, 211

A Brief History of DNA Evidence, 211

Relevant Expertise, 214

Variation in Human DNA and Its Detection, 216

What Are DNA, Chromosomes, Genes, Proteins, and RNAs?, 216

The DNA Molecule, 216

Chromosomes, 216

Genes and Gene Products, 217

What are genes, proteins, and RNAs?, 217

How do cells manufacture proteins and RNAs?, 219

Alleles and Loci of Genes, 219

Sexual Reproduction and the Genome, 220

What Are DNA Polymorphisms and How Are They Detected?, 222

VNTRs and RFLP Testing, 222

PCR Amplification, 223

STRs and Capillary Electrophoresis, 224

How STRs differ from VNTRs, 224

Capillary electrophoresis and fluorescence, 226

Miniaturization for rapid capillary electrophoresis, 230

Sequence-Specific Probes and Microarrays, 233

Sequencing and SNPs, 234

Sequencing methods, 235

The range of forensic applications, 238

SNPs as identification loci, 239

- Summary, 241
- What Establishes That a Genetic System Is Valid for Identification?, 241
- Sample Collection and Laboratory Performance, 243
 - Sample Collection and Contamination, 243
 - Did the Sample Contain Enough DNA?, 244
 - Was the Sample of Sufficient Quality?, 246
 - Laboratory Performance, 248
 - What Quality Control and Assurance Measures Are in Place?, 248
 - Documentation, 249
 - Validation, 250
 - Proficiency testing, 251
 - How Are Samples Created and Handled?, 253
- Inference, Statistics, and Population Genetics in Human Nuclear DNA Testing, 256
 - What Constitutes a Match? An Exclusion?, 256
 - What Hypotheses Can Be Formulated About the Source?, 257
 - Can the Match Be Attributed to Laboratory Error?, 258
 - Could a Close Relative Be the Source?, 259
 - Could an Unrelated Person Be the Source?, 260
 - Frequencies, Probabilities, and Prejudice, 263
 - Are Frequencies or Probabilities Prejudicial Because They Are So Small?, 264
 - Are Frequencies or Probabilities Prejudicial Because They Might Be Transposed?, 265
 - Are Random-Match Probabilities That Are Smaller Than False-Positive Error Probabilities Irrelevant or Prejudicial?, 267
 - “Rarity,” Source, and Uniqueness Testimony, 269
- Special Issues in Human DNA Testing, 272
 - Y Chromosomes, 272
 - Genetic Principles, 272
 - Forensic Value, 273
 - Analytical Methods and Categorical Matches, 274
 - Population Genetics and Statistics, 275
 - Haplotype frequency or probability estimates, 276
 - Likelihood ratios (LRs), 278
 - Mitochondrial DNA, 279
 - Mitochondria and Their Genomes, 279
 - Forensic Value, 280
 - Analytical Methods and Categorical Matches, 281
 - Population Genetics and Statistics for Mitochondrial DNA, 283
 - Population databases, 283
 - Heteroplasmy, 284
 - Emerging Questions for Courts, 286
 - Mixtures, 288

Reference Guide on Human DNA Identification Evidence

- What Are DNA Mixtures?, 288
- Have Strategies for Simplifying Mixtures Interpretation Been Pursued?, 290
- What Is Mixture Deconvolution?, 291
- What Statistical Assessment of the Comparison to a Defendant’s Profile Has Been Conducted?, 294
 - Probability of inclusion or exclusion, 295
 - Combined probability of inclusion (CPI) and exclusion (CPE), 295
 - Random-match probability (RMP), 296
 - Likelihood ratio (LR), 297
 - Definition of the likelihood ratio, 297
 - Binary genotyping, 300
 - Probabilistic genotyping software (PGS), 301
 - Validation of PGS programs, 304
 - Presentation of LRs, 309
- DNA and RNA Typing of Bodily Fluids or Tissues, 312
- Twins, 313
- Investigative DNA Analysis, 317
 - Offender and Suspect Database Searches, 317
 - Law-enforcement databases and databanks, 317
 - Probative value of database hits, 320
 - The statistical analyses of adjustment, 320
 - Judicial opinions on adjustment, 322
 - Kinship trawling (“familial searching”), 323
 - All-pairs matching within a database to verify estimated random-match probabilities, 325
 - Investigative Genetic Genealogy (IGG), 327
 - How does direct-to-consumer genetic genealogy work?, 328
 - How does IGG work?, 330
 - From the forensic sample to the SNP genotype, 331
 - From the SNP genotype to possible relatives, 332
 - Does IGG violate the Fourth Amendment?, 332
 - Are parts of IGG a search or seizure?, 333
 - Warrants and probable cause?, 336
 - How should IGG be regulated?, 337
 - Inferring Traits, 339
 - Biogeographic ancestry testing (BGAT), 340
 - Forensic DNA phenotyping (FDP), 341
- Glossary of Terms, 345
- References on DNA, 360
 - Genetics and Genomics, 360
 - Forensic Genetics and Genomics, 360

FIGURES

- 1. Sketch of a small part of a double-stranded DNA molecule. Nucleotide bases are held together by weak bonds. A pairs with T; C pairs with G, 217
- 2. Three alleles of the D16S539 STR, 225
- 3. Sketch of an electropherogram for two D16S539 alleles. 227
- 4. Alleles of 15 STR loci and the amelogenin sex-typing test from the AmpFlSTR Identifiler kit, 229
- 5. Electropherogram for nine STR loci of the victim’s DNA in *People v. Pizarro*. 230
- 6. Electropherogram in *People v. Pizarro* that can be interpreted as a mixture of DNA from the victim and the defendant, 289

TABLE

- 1. Hypotheses That Might Explain a Match Between Defendant’s DNA and DNA at a Crime Scene, 258

Introduction

Deoxyribonucleic acid, or DNA, is a molecule that encodes the genetic information in all living organisms. Its chemical structure was elucidated in 1953. More than 30 years later, samples of human DNA began to be used in the criminal justice system, primarily in cases of rape or murder. The evidence has been the subject of extensive scrutiny by lawyers, judges, and the scientific community. DNA evidence is admissible in all jurisdictions, but there are many types of forensic DNA analysis, and still more are being developed. Questions of admissibility and weight of DNA evidence arise as advancing methods of biochemical and statistical analysis, along with novel applications of established methods, are introduced. Moreover, questions about the appropriate balance between law-enforcement goals and individual privacy and security also emerge when law-enforcement officials collect and analyze human DNA samples. This reference guide addresses technical issues that are important when considering the admissibility of and weight to be accorded analyses of human DNA and the related molecule RNA (ribonucleic acid) in trials. In addition, this guide describes ways in which human DNA assists in criminal investigations, and it identifies a number of the legal and policy questions raised by investigative genetics.

A Brief History of DNA Evidence

“DNA evidence” refers to the results of chemical or physical tests that directly reveal differences in DNA molecules found in organisms as diverse as bacteria, plants, and animals.¹ The technology for establishing the identity of individuals from the DNA in blood, semen, and other bodily fluids became available to law-enforcement agencies in the mid to late 1980s.² The judicial reception of DNA evidence can be divided into at least six phases.³ The first phase was one of rapid acceptance. Initial praise for RFLP (restriction fragment length polymorphism)

1. Differences in DNA also can be revealed by differences in the proteins that are made according to the “instructions” in a DNA molecule. Blood-group factors, serum enzymes and proteins, and tissue types all reveal information about the DNA that codes for these chemical structures. Such immunogenetic testing, including the application of antibodies to detect cell-surface antigens, predates the more direct DNA testing that is the subject of this guide. On the nature and admissibility of such testing, see, for example, David H. Kaye, *The Double Helix and the Law of Evidence* 5–19 (2010); 1 *McCormick on Evidence* § 205(B) (Robert Mosteller ed., 8th ed. 2020). See generally Liesa L. Richter & Daniel J. Capra, *Reference Guide on the Admissibility of Scientific Evidence*, in this manual.

2. The first reported appellate opinion is *Andrews v. State*, 533 So. 2d 841 (Fla. Dist. Ct. App. 1988).

3. The description that follows is adapted and updated from 1 *McCormick on Evidence*, *supra* note 1, § 205(B). Beyond these six phases for the principal methods of human DNA profiling encountered in courts, the supplementation of STR profiling with other genetic markers for identification related to single-nucleotide differences is beginning.

testing in homicide, rape, paternity, and other cases was effusive.⁴ Expert testimony rarely was countered, and courts readily admitted DNA evidence.

In a second wave of cases, however, defendants pointed to problems at two levels—controlling the experimental conditions of the analysis and interpreting the results. Concerted attacks by defense experts with impressive credentials led to the rejection of specific proffers on the grounds that the testing was not sufficiently rigorous.⁵

A different attack on DNA profiling that began in cases during this period led to a third wave of cases in which many courts held that estimates of the probability of a coincidentally matching DNA profile were inadmissible. These estimates relied on a simple population-genetics model for the frequencies of DNA profiles, and some prominent scientists claimed that the applicability of the mathematical model had not been adequately verified. A heated debate on this point spilled over from courthouses to scientific journals and convinced the supreme courts of several states that general acceptance was lacking. A 1992 report of the National Academy of Sciences (NAS) proposed a more “conservative” computational method as a compromise,⁶ and this seemed to undermine the claim of scientific acceptance of the procedure that was in general use.

An outpouring of critiques of the report led to the formation of a second NAS committee. This panel concluded in 1996 that the usual method of estimating frequencies in broad population groups generally was sound, and it proposed improvements and additional procedures for estimating frequencies in subgroups.⁷ In the corresponding fourth phase of judicial scrutiny of DNA evidence, the courts almost invariably returned to the earlier view that the statistics associated with DNA profiling are generally accepted and scientifically valid.

4. *E.g.*, *People v. Wesley*, 140 Misc. 2d 306, 533 N.Y.S.2d 643, 644 (Alb. Cnty. Ct. 1988) (“the single greatest advance in the ‘search for truth’ . . . since the advent of cross-examination”). DNA testing also quickly became powerful evidence of kinship in child support, immigration, and other cases involving genetic parentage or other relationships within a family.

5. Moreover, a minority of courts, perhaps concerned that DNA evidence might be conclusive in the minds of jurors, added a “third prong” to the general-acceptance standard of *Frye v. United States*, 293 F. 1013 (D.C. Cir. 1923). This augmented *Frye* test requires not only proof of the general acceptance of the ability of science to produce the type of results offered in court, but also of the proper application of an approved method on the particular occasion. For commentary, see David L. Faigman et al., *Modern Scientific Evidence* § 30:9 n.5 (2022–2023 ed.) (citing articles); David H. Kaye et al., *The New Wigmore, A Treatise on Evidence: Expert Evidence* § 7.3.3(a)(2) (3d ed. 2021) (criticizing the approach); Joseph G. Petrosinelli, Comment, *The Admissibility of DNA Typing: A New Methodology*, 79 Geo. L. J. 313, 327–31 (1990) (similar criticism); *cf.* Ming W. Chin et al., *Forensic DNA Evidence: Science and the Law* § 11:6 (2022) (describing California’s “limited third-prong hearing”).

6. National Research Council, *DNA Technology in Forensic Science* (1992) [hereinafter NRC I].

7. National Research Council, *The Evaluation of Forensic DNA Evidence* (1996) [hereinafter NRC II].

In the fifth phase of the judicial evaluation of DNA evidence, results obtained with PCR-based methods⁸ entered the courtroom. The opinions were practically unanimous in holding that the PCR-based procedures rested on a solid scientific foundation and were generally accepted in the scientific community. Before long, forensic scientists settled on the use primarily of one type of DNA variation—known as short tandem repeats, or STRs—to include or exclude individuals as the source of crime-scene DNA. Investigative databases of STR profiles from men, women, and children convicted (and sometimes only arrested for) specified crimes grew by leaps and bounds.⁹ Matching these recorded profiles to STR profiles of crime-scene samples generated seemingly magical investigative leads.

The sixth phase of judicial scrutiny is still in progress. With the extension of STR profiling to minute quantities of DNA (low template, or LT-DNA), often from more than one individual, laboratories have turned to computerized methods to infer which individual profiles may be giving rise to the patterns of STRs and which features in the data are the result of instrumental error or limitations. Both the modifications for generating data from LT-DNA and the “probabilistic genotyping software” have been the subject of admissibility hearings.

Less frequently, other genetic systems to establish the identity of the source of trace DNA evidence have been applied to generate investigative leads. The highly publicized technique of trawling commercial or other privately maintained databases established for genealogical research, to discover individuals who might be related to the source, is discussed in the section titled “Investigative Genetic Genealogy” below. Inferring visible traits and biogeographic ancestry from DNA found at crime scenes is described in the section titled “Forensic DNA Phenotyping” below. These investigative procedures rely on newer methods for analyzing DNA that are just beginning to be evaluated in court.

Throughout these phases, DNA tests also exonerated an increasing number of people who had been convicted of capital and other crimes, posing serious questions about the adequacy of the legal procedures for obtaining postconviction DNA testing and for overturning factually erroneous convictions.¹⁰ The value of

8. PCR (polymerase chain reaction) is described in the section titled “Chromosomes” below.

9. See *Maryland v. King*, 569 U.S. 435 (2013) (upholding compulsory DNA sampling as part of the booking process for custodial arrests); section titled “Law-enforcement databases and databanks” below. Initially, the law-enforcement databases housed records of restriction-fragment-length polymorphisms arising from variable numbers of tandem repeats (RFLP-VNTR profiles). FBI, Frequently Asked Questions on CODIS and NDIS, <https://www.fbi.gov/how-we-can-help-you/dna-fingerprint-act-of-2005-expungement-policy/codis-and-ndis-fact-sheet>, last visited Sept. 8, 2024 (“The National DNA Index no longer searches DNA data developed using restriction fragment length polymorphism (RFLP) technology.”).

10. See, e.g., *Osborne v. Dist. Atty’s Office for Third Jud. Dist.*, 557 U.S. 52 (2009) (narrowly rejecting a convicted offender’s claim of a due process right to DNA testing at his expense, enforceable under 42 U.S.C. § 1983, to establish that he is probably innocent of the crime for which he was convicted after a fair trial, when (1) the convicted offender did not seek extensive DNA testing before trial even though it was available, (2) he had other opportunities to prove his innocence after

DNA evidence in solving older crimes also prompted extensions of some statutes of limitations.¹¹

In sum, in little more than a decade, forensic DNA typing made the transition from a novel set of methods for identification to a relatively mature and well-studied forensic technology. However, one should not lump all forms of DNA identification together. New techniques and applications continue to emerge. Before admitting DNA evidence, courts normally inquire into the biological principles and knowledge that would justify inferences from new or different technologies or applications. As a result, this guide describes not only the predominant STR technology and earlier methods that were more controversial, but also newer analytical techniques that can be used in human forensic DNA identification.

Relevant Expertise

Human DNA identification can involve testimony about laboratory findings, about the statistical interpretation of those findings, and about the underlying principles of molecular biology and genetics. Consequently, expertise in several fields might be required to establish the admissibility of the evidence or to explain it adequately to the trier of fact. The expert who is qualified to testify about laboratory techniques might not be qualified to testify about molecular biology, to make estimates of population frequencies, or to establish that an estimation procedure is valid.¹²

Trial judges ordinarily are accorded great discretion in deciding when a witness is qualified to testify as an expert, and these decisions depend on the

a final conviction based on substantial evidence against him, (3) he had no new evidence of innocence (only the hope that more extensive DNA testing than that done before the trial would exonerate him), and (4) even a finding that he was not the source of the DNA would not conclusively demonstrate his innocence); *Skinner v. Switzer*, 562 U.S. 521 (2011); Nat'l Comm'n on the Future of DNA Evidence, *Postconviction DNA Testing: Recommendations for Handling Requests* (1999); Brandon L. Garrett, *Judging Innocence*, 108 Colum. L. Rev. 55 (2008); Brandon L. Garrett, *Claiming Innocence*, 92 Minn. L. Rev. 1629 (2008).

11. See, e.g., Veronica Valdivieso, Note, *DNA Warrants: A Panacea for Old, Cold Rape Cases?* 90 Geo. L.J. 1009 (2002).

12. See David H. Kaye & Hal S. Stern, *Reference Guide on Statistics and Research Methods*, section titled "Relevant Expertise," in this manual. Nonetheless, if previous cases establish that the testing and estimation procedures are legally acceptable, and if the computations are essentially mechanical, then highly specialized statistical expertise might not be essential. Reasonable estimates of DNA characteristics in major population groups can be obtained from standard references, and many quantitatively literate experts could use the appropriate formulae to compute the relevant profile frequencies or probabilities. NRC II, *supra* note 7, at 170. Limitations in the knowledge of a technician who applies a generally accepted statistical procedure can be explored on cross-examination. E.g., *Roberson v. State*, 16 S.W.3d 156, 168 (Tex. Crim. App. 2000).

background of each witness. Courts have noted the lack of familiarity of academic experts—who have done respected work in other fields—with the scientific literature on forensic DNA typing and on the extent to which their research or teaching lies in other areas.¹³ Although such concerns may affect the persuasiveness of particular testimony, they rarely result in exclusion on the grounds that the witness simply is not qualified as an expert on a particular topic.¹⁴

Because forensic DNA analysis generally draws on methods and technology developed for research and applications in biological and medical science, the scientific literature on the underpinnings of most forms of DNA evidence, both established and emerging, tends to be extensive. By studying the scientific publications, or perhaps by appointing a special master or expert adviser to assimilate this material,¹⁵ a court can ascertain where a party's expert falls within the spectrum of scientific opinion. Furthermore, an expert appointed by the court under Federal Rule of Evidence 706 could testify about the scientific literature generally or even about the strengths or weaknesses of the particular arguments advanced by the parties.¹⁶

Given the diversity of forensic questions to which DNA testing might be applied, it is not feasible to list the specific scientific expertise appropriate to all applications. Some applications span many fields. Consider the range of knowledge required to assess the value of DNA analyses of a novel application such as

13. *E.g.*, *State v. Copeland*, 922 P.2d 1304, 1318 n.5 (Wash. 1996) (noting that defendant's statistical expert "was also unfamiliar with publications in the area," including studies by "a leading expert in the field" whom he thought was "a 'guy in a lab somewhere'").

14. *E.g.*, *United States v. Pritchard*, 993 F. Supp. 2d 1203, 1208–09 (C.D. Cal. 2014) (laboratory analyst "easily meets the standard for qualification" regarding "statistical analysis testimony" because of her academic education and laboratory and courtroom experience, as supplemented by "special training in statistics" at three or four short courses between 8 and 24 hours in length). *But see* *Allen v. State*, 62 So. 3d 1199 (Fla. Dist. Ct. App. 2011) (remanding for "a limited evidentiary hearing on the qualifications of the state's expert to testify as to the statistical significance of the DNA profile matches" where witness had taken statistics in college and received on-site training, but the state did not meet its burden of showing the analyst's proficiency and general acceptance of the statistical methodology); *Gibson v. State*, 915 So. 2d 199 (Fla. Dist. Ct. App. 2005) (although the analyst had taken courses in statistics and had testified that she was following the procedure in the 1996 NRC report, the court of appeals "remanded for a limited evidentiary hearing to determine whether the expert had sufficient knowledge of the authoritative sources to present the statistical evidence"). More case law is collected in George L. Blum, Annotation, *Qualification as Expert to Testify as to Findings or Results of Scientific Test Concerning DNA Matching*, 38 A.L.R. 6th 439 (2008).

15. *See, e.g.*, *Gen. Elec. Co. v. Joiner*, 522 U.S. 136, 149–50 (1997) (Breyer, J., concurring) (discussing the use of "special masters and specially trained law clerks" in science-related cases); *Ass'n of Mex.-Am. Educators v. State of Cal.*, 231 F.3d 572, 590 (9th Cir. 2000) ("[i]n those rare cases in which outside technical expertise would be helpful to a district court, the court may appoint a technical advisor"); *United States v. Lewis*, 442 F. Supp. 3d 1122 (D. Minn. 2020) (relying on the report of a special master on the validation and performance of "probabilistic genotyping" software used for the analysis of complex DNA mixtures).

16. *See generally* *Kaye et al.*, *supra* note 5, § 12.2; *Faigman et al.*, *supra* note 5, §§ 1:37 to 1:39.

whole genome sequencing (see section titled “Sequencing and SNPs” below) to establish that a semen sample in a rape case originated from one identical twin rather than the other (see section titled “Twins” below). Expertise in molecular genetics, embryology, biotechnology, bioinformatics, and statistics might be necessary to evaluate the premises and execution of the analysis. Generally, when samples come from crime scenes, the expertise and experience of forensic scientists can be crucial. Just as highly focused specialists may be unaware of aspects of an application outside their field of expertise, so too scientists who have not previously dealt with forensic samples can be unaware of case-specific factors that can confound the interpretation of test results.

Variation in Human DNA and Its Detection

What Are DNA, Chromosomes, Genes, Proteins, and RNAs?

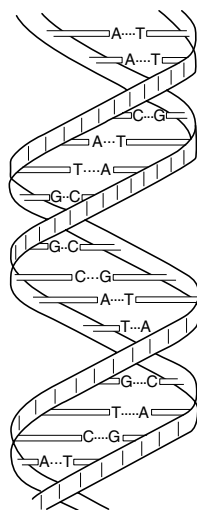
The DNA Molecule

DNA, RNA, and protein molecules are all polymers—large molecules composed of a chain of subunits. The subunits of DNA include four chemical structures known as nucleotide bases. Their names—adenine, thymine, guanine, and cytosine—usually are abbreviated as A, T, G, and C, respectively. The physical structure of DNA is often described as a double helix because the molecule has two spiraling strands connected to each other by weak bonds between the nucleotide bases. As shown in Figure 1, A pairs only with T, and G pairs only with C. Thus, the order of the single bases on either strand reveals the order of the pairs from one end of the molecule to the other, and the DNA molecule represents a long sequence of As, Ts, Gs, and Cs.

Chromosomes

Most human DNA is tightly packed into structures known as chromosomes, which come in different sizes and are located in the nuclei of cells. The chromosomes are numbered, in descending order of size, 1 through 22, with the remaining chromosome being an X or a much smaller Y. Because one of each of these 23 chromosomes is inherited from each parent, the cell’s nucleus normally houses 46 chromosomes in all (see the section titled “Sexual Reproduction and the Genome” below). If the bases are like letters, then each chromosome is like a book written in this four-letter alphabet, and the nucleus is like a bookshelf in the

Figure 1. Sketch of a small part of a double-stranded DNA molecule. Nucleotide bases are held together by weak bonds. A pairs with T; C pairs with G.



interior of the cell. All the cells in one individual contain identical copies of the same collection of books.¹⁷ The sequence of the As, Ts, Gs, and Cs that constitutes the “text” in these chromosomes is referred to as the individual’s nuclear genome.

All told, the genome has more than three billion “letters” (As, Ts, Gs, and Cs). If these letters were printed in books, the resulting pile would be as high as the Washington Monument. About 99.9% of the genome is identical between any two individuals. This similarity is not really surprising—it accounts for the common features that make modern humans an identifiable species (and for features that we share with many other species as well). The remaining 0.1% is particular to an individual. This variation makes each person (other than identical twins) genetically unique. This small percentage may not sound like a lot, but it adds up to more than three million sites for variation among individuals.

Genes and Gene Products

What are genes, proteins, and RNAs?

Variations in the sequence of base pairs within human populations occur both within the genes and in the regions between the genes. A gene can be defined as

17. Occasional mutations occur, usually as a result of mistakes in the duplication of chromosomes during cell division. See *infra* note 25.

a segment of DNA, usually from 1,000 to 10,000 base pairs long, that “encodes” a protein or an RNA molecule. For example, a large gene named GC (for group-specific component) encodes a protein that binds to vitamin D and is found in blood plasma. In many people, the following tiny sequence appears within that gene:

G C A A A A T T G C C T G A T G C C A C A C C C A A G G A A C T G
G C A.

Unlike double-stranded DNA, RNA is a single helical strand that has a different nucleotide (U, which is short for uracil) in place of the T in the DNA helix. As indicated below, RNA plays an integral part in gene expression.

Proteins are composed of units called amino acids. Twenty different amino acids exist in proteins, and hundreds to thousands of them are attached to each other in long chains with more complex shapes than the single helix of RNA or the double helix of DNA. Proteins perform all sorts of functions in the body and thus produce observable characteristics. For example, the group-specific component protein referred to above (and designated Gc) grabs onto vitamin D and circulates it in the body. It also plays a role in the immune system.

The order of the building blocks of this protein can be slightly different in different individuals. Forensic scientists used these differences before DNA testing was available to exclude or include suspects or defendants as possible sources of blood or semen stains. The cell produces specific proteins and RNAs that correspond to the order of the bases (the “letters”) in the coding part (exons) of a gene. The sequence in which the protein’s amino acids are arranged corresponds to the sequence of base pairs within a gene. A sequence of three base pairs (a codon) specifies a particular 1 of the 20 possible amino acids in the protein. The mapping of various sequences of three nucleotide bases to a particular amino acid is the genetic code. About 1.5% of the human genome codes for the amino-acid sequences.

Human genes also contain noncoding sequences that regulate the cell type in which a protein will be synthesized and how much protein will be produced. Many genes contain interspersed noncoding, nonregulatory sequences that no longer participate in protein synthesis. These sequences, called introns, constitute about 23% of the base pairs within human genes. In terms of the metaphor of DNA as text, the gene is like an important paragraph in the book, often with some gibberish in it.¹⁸

18. The idea of a gene as a block of DNA (some of which is coding, some of which is regulatory, and some of which is functionless) is an oversimplification (*see, e.g.,* Petter Portin & Adam Wilkins, *The Evolving Definition of the Term “Gene,”* 205 *Genetics* 1353 (2017)), but it is useful enough here.

How do cells manufacture proteins and RNAs?

Protein-coding genes express their sequence-specific information by an intricate process.¹⁹ Through a series of biochemical reactions, the base pairs of the exons of the gene are transcribed into an RNA molecule. The noncoding parts (the introns) are not represented in this messenger RNA (mRNA) transcript. The transcribed mRNA makes its way to microscopic molecular factories, where it is translated into a protein.²⁰ By combining the amino acids in different orders, a virtually unlimited variety of proteins can be constructed. Other genes contain DNA sequences that are transcribed into RNAs that perform other functions. Their final products are nonprotein-coding RNAs (ncRNAs).²¹

Alleles and Loci of Genes

The GC gene mentioned above always is located at the same position, on chromosome 4. But the short sequence listed as occurring within that gene is not the same for every individual. A locus where almost all humans have the same DNA sequence is called monomorphic (“of one form”). A locus where the DNA sequence varies among significant numbers of individuals—more than 1% or so of the population possesses the variant—is called polymorphic (“of many forms”). The alternative forms are called alleles. The GC locus, for example, is

19. The description here of how genes express proteins and RNAs is adapted, in part, from a more complete explanation in the Brief of Genetics, Genomics and Forensic Science Researchers as Amici Curiae in Support of Neither Party in *Maryland v. King*, 569 U.S. 435 (2013) (reprinted in Henry T. Greely & David H. Kaye, *A Brief of Genetics, Genomics and Forensic Science Researchers in Maryland v. King*, 53 *Jurimetrics J.* 43 (2013)) [hereinafter Amici Brief].

20. Transfer RNA brings the building blocks of proteins (the amino acids harvested from food) to these structures. There, they are joined in the order dictated by the mRNA transcript.

21. The ncRNAs include RNAs that do the work of translation and RNAs involved in regulating expression of dozens or even hundreds of protein-coding genes. Furthermore, much shorter RNAs regulate transcription and translation. Thus, it is widely recognized that the genome is abuzz with transcription-to-RNA activity and other events that interact in the expression of the protein-coding DNA. The discoveries of these processes generated speculation in the press and some judicial opinions that *all* DNA sequences significantly affect development and health. The *King* amici maintained that

This is not true—even for sequences that are transcribed. Not every biochemical event along the DNA has a measurable impact on health. First, some short ncRNA transcripts are just “noise.” They are degraded quickly. Second, intronic parts of the pre-mRNA transcripts normally are removed. Third, even if one transcript could regulate expression, other transcripts may do the same job. Fourth, even if the stable transcript does affect some trait, the effect may be so minor in the context of other genetic and environmental influences as to be of no meaningful predictive or diagnostic value. Finally, even if a stable transcript has a dramatic effect on a trait, that trait may be unrelated to disease status. Consequently, regarding every bit of the genome as if it were a medical record is unwarranted.

Amici Brief, *supra* note 19, at 11.

polymorphic. The sequence has three common alleles that result from substitutions in a base at a given point. Where an A appears in one allele, there is a C in another. The third allele has the A, but at another point a G is swapped for a T. These changes are called single nucleotide polymorphisms (SNPs, pronounced “snips”).

If a gene is like a paragraph in a book, a SNP is a change in a letter somewhere within that paragraph (a substitution, a deletion, or an insertion), and the two versions of the gene that result from this slight change are the alleles. An individual who inherits the same allele from both parents is called a homozygote. An individual with distinct alleles is a heterozygote.

DNA sequences used for forensic analysis usually are not inside the coding parts of genes. They lie in the vast regions between genes (about 75% of the genome is extragenic) or in the introns. These extra- and intragenic regions of DNA have been found to contain considerable sequence variation, which makes them particularly useful in distinguishing individuals. Although the terms “locus,” “allele,” “homozygous,” and “heterozygous” were developed to describe genes, the nomenclature has been carried over to describe all DNA variation—coding and noncoding alike. Both types are inherited from mother and father in the same fashion, as discussed in the next subsection.

Sexual Reproduction and the Genome

The process that gives rise to the diversity of alleles starts with the production of special sex cells—sperm cells in males and egg cells in females. All the nucleated cells in the body other than sperm and egg cells contain two versions of each of the 23 chromosomes—two copies of chromosome 1, two copies of chromosome 2, and so on, for a total of 46 chromosomes. The 22 numbered chromosomes are called autosomes. Cells in females contain two X chromosomes, and cells in males contain one X and one Y chromosome.²² An egg cell, however, contains only 23 chromosomes—one chromosome 1, one chromosome 2, . . . , one chromosome 22, and one X chromosome—each selected at random from the woman’s full complement of 23 chromosome pairs. Thus, each egg carries half the genetic information present in the mother’s 23 chromosome pairs—a “haploid” genome. Because the assortment of the chromosomes is random, each egg carries a different complement of genetic information. Further randomness comes from a process known as crossing over, in which segments of each of the mother’s two paired chromosomes exchange places in producing the single chromosome from the pair that ends up in

22. A small fraction of people are born with extra chromosomes or a missing one. The most common such condition is Down syndrome (an extra chromosome 21), which occurs in about 1 in every 700 babies born. Nat’l Center on Birth Defects & Developmental Disabilities, Centers for Disease Control and Prevention, Data and Statistics on Down Syndrome (Dec. 16, 2022), <https://perma.cc/VB5W-XM5D>.

the egg. The same situation exists with sperm cells. Each sperm cell contains a single copy of each of the 23 chromosomes (with crossing over) selected at random from a man's 23 pairs.²³ Fertilization of an egg by a sperm therefore restores the full number of 46 chromosomes, with the 46 chromosomes in the fertilized egg—a “diploid” genome—that is a new combination of those in the mother and father. This recombination of genetic information in sexual reproduction is the main source of human genetic diversity.

During pregnancy, the fertilized cell divides to form two cells, each of which has an identical copy of the 46 chromosomes. The two then divide to form four, the four form eight, and so on. As gestation proceeds, various cells specialize (differentiate) to form different tissues and organs. Although cell differentiation yields many different kinds of cells, the process of cell division usually results in a line of cells having the same genomic complement as the cell that divided. Thus, aside from occasional mutations occurring after fertilization,²⁴ each of the approximately 100 trillion cells in the adult human body has the same DNA as was present in the original 23 pairs of chromosomes from the fertilized egg, one member of each pair having come from the mother and one from the father.²⁵

23. The man's two sex chromosomes (an X and a Y) are so different that they do not undergo crossing over during the formation of a sperm cell (except in small “pseudo-autosomal regions”—regions at one or both ends of the pair).

24. Mutations in a parent's sex cell (an egg or sperm) that occur before conception will be incorporated into the genome of the progeny by this repeated copying process. Such pre-fertilization mutations could complicate parentage or other relationship testing, but they do not matter when testing to see whether two samples originated from the same progeny—for example, a suspect in a criminal case.

25. Mutations can occur during embryonic development or later. The most important source of these changes is a mistake in the DNA copying process during cell division. These are low probability events, but given the immense number of cell divisions that occur during the life of an organism, some are to be expected. Thus, every human being may well be a genetic mosaic to some degree. The mutation rate seems to vary across sites within the genome and between different cell types. Some mutations result in cancers—cells whose proliferation is not inhibited by normal mechanisms. Cristian Tomasetti et al., *Stem Cell Divisions, Somatic Mutations, Cancer Etiology, and Cancer Prevention*, 355 *Science* 1330 (2017), <https://doi.org/10.1126/science.aaf9011>. New cell lineages arising after birth (cancerous or otherwise) do not normally confuse forensic DNA comparisons because the vast majority of them would not occur at the loci used in identity testing (see section titled “What Are DNA Polymorphisms and How Are They Detected?” below), and even if they did, the fraction of the mutated cells in most forensic DNA samples would be so small that only the unmutated alleles would be detected. It is also possible that a mutation would occur “so early in embryonic development that DNA in eggs or sperm might differ from that in blood [or saliva] from the same person.” NRC II, *supra* note 7, at 64 n.7. The NRC report characterizes this mosaicism as “a remote possibility.” *Id.* Whether or not this is accurate, as with mutations occurring later in life, the embryonic mutations resulting in the divergence between tissues would have to change the specific loci used in forensic identity testing, which is additionally improbable.

What Are DNA Polymorphisms and How Are They Detected?

By determining which alleles are present at strategically chosen locations on a chromosome (loci), the forensic scientist ascertains the DNA profile, or genotype, of an individual (at those loci). Although the differences among the alleles arise from alternations in the order of nucleotide base pairs (the A, T, G, C letters), DNA typing for ascertaining identity does not require “reading” the full genome. Here we outline the major types of polymorphisms that have been used in identity testing and the methods for detecting them.

VNTRs and RFLP Testing

The first polymorphisms to find widespread use in identity testing result from the presence of a variable number of tandem repeats (VNTRs) at a locus. Being the subject of most of the court opinions on the admissibility of DNA evidence in the late 1980s and early 1990s, they paved the way for the methods now in use. These VNTRs are one type of repetitive DNA. They are like a musical score in which the same melody is played over and over without interruption. The core unit (the melody) is a particular short DNA sequence, and the number of times it is repeated varies from one person to another. The first VNTRs to be used in genetic and forensic testing had core repeat sequences of 7–35 or more base pairs.

In forensic VNTR testing, enzymes (known as restriction enzymes) were used to cut the DNA molecule both before and after the VNTR sequence. A small number of repeats in the VNTR region gives rise to a small restriction fragment, and a large number of repeats yields a large fragment. A substantial quantity of DNA is required to give a detectable number of VNTR fragments with this procedure. After the restriction fragments are sorted by size with a process known as gel electrophoresis,²⁶ the fragments with particular VNTRs are detected by applying a “probe” that binds when it encounters the repeated core sequence. A probe is a short, synthesized fragment of single-stranded DNA with base pairs arranged to complement the core sequence. For example, if the core’s sequence is AGTTC-GCGTGACTAT, the probe’s sequence would be TCAAGCGCAGGATA. Because A bonds to T and G to C (and vice versa; see Figure 1), when the probe comes in contact with the restriction fragment, it sticks to it. A radioactive

26. Electrophoresis separates DNA, RNA, or protein molecules according to their size. An electric current drags the molecules through a gel or other matrix. Smaller molecules move through the matrix more quickly than larger molecules. Molecules of known sizes (“standards”) are run through the gel at the same time as the molecules from the sample. Comparing their positions at the end of the run to those of the separated molecules indicates the sizes of the latter.

or fluorescent molecule attached to the probe provides a way to mark the VNTR fragment. Many court opinions refer to this process as RFLP testing.²⁷

PCR Amplification

Most modern forensic-science methods of DNA analysis take advantage of a chemical reaction called PCR (for polymerase chain reaction). The PCR process enables laboratories to make many copies of small DNA fragments.²⁸ Other procedures then can be applied to analyze the vastly amplified quantity of DNA.

PCR can be applied to the double-stranded DNA segments extracted and purified from a forensic sample as follows: First, the purified DNA is separated into two strands by heating it to near the boiling point of water.²⁹ Second, the single strands are cooled, and primers—synthesized fragments of single-stranded DNA, usually between 15 and 30 nucleotides long—attach themselves to the points at which the copying will start and stop.³⁰ Finally, the soup containing the annealed DNA strands, an enzyme called DNA polymerase, and lots of the four nucleotide building blocks are warmed. On a DNA strand, the polymerase inserts the complementary bases one at a time, building a new strand bound to the original template and thus replicating part of the DNA strand that was separated from its partner in the first step. The same replication occurs with the separated partner as the template.³¹ The result is two identical double-stranded DNA segments, one made from each strand of the original DNA. The three-step cycle is repeated, usually 20 to 35 times in automated machines known as thermal cyclers. Ideally, starting with a single double-stranded molecule, the first cycle results in two double-stranded DNA segments; the second cycle produces four; the third, eight; and so on, until there are millions of copies of the DNA sequences of interest.³²

Care must be taken to achieve the appropriate chemical conditions. Furthermore, because even small amounts of stray DNA could be amplified along with the DNA of interest, quality-assurance steps must be taken to reduce contamination

27. It would be clearer to call it RFLP-VNTR testing, because the fragments being measured contain the VNTRs rather than some simpler polymorphisms that were used in genetic research and disease testing. A more detailed exposition of the steps in RFLP-VNTR profiling (including gel electrophoresis, Southern blotting, and autoradiography) can be found in the first edition of this guide (Reference Manual on Scientific Evidence (1st ed. 1994)) and in many judicial opinions circa 1990.

28. Most VNTRs are too long for PCR to work efficiently.

29. This “denaturing” takes about a minute. Chemical and other physical methods for denaturing also can be used.

30. By judiciously constructing the primers to bracket the sequences of interest, the amplified DNA will consist almost entirely of the in-between sequences. The rest of the genome will not be copied. Annealing the primers takes about 45 seconds.

31. This extension step for both templates takes about two minutes.

32. In practice, there is some inefficiency in the doubling process, but the yield from a 30-cycle amplification is generally about 1 million to 10 million copies. NRC II, *supra* note 7, at 69–70.

of the sample.³³ A laboratory should be able to demonstrate that it can amplify targeted sequences faithfully with the equipment and reagents that it uses and that it has taken suitable precautions to minimize or detect contamination from extraneous DNA.³⁴ With small samples, it is possible that some alleles will be amplified and others missed (preferential amplification, discussed below in the section titled “Did the Sample Contain Enough DNA?”). Mutations within a population giving rise to variants in the region of a primer can prevent the amplification of the allele downstream of the primer, giving rise to null alleles.³⁵

STRs and Capillary Electrophoresis

How STRs differ from VNTRs

Although RFLP-VNTR profiling is highly discriminating,³⁶ it has several drawbacks. Not only does it require a substantial sample of DNA-bearing material, but it is time-consuming and does not measure the fragment lengths to the nearest number of repeats.³⁷ Consequently, forensic scientists moved from VNTRs to another form of repetitive DNA known as short tandem repeats

33. In some instances, interpretation of results may be valuable notwithstanding known contaminating DNA. For example, it may be possible to exclude an individual as a viable source. See Human Forensic Biology Subcomm., Organization of Scientific Area Committees for Forensic Science, Standard for Interpreting, Comparing and Reporting DNA Test Results Associated with Failed Controls and Contamination Events, June 1, 2021 (OSAC 2020-S-0004).

34. Forensic-science organizations have proposed guidelines. See, e.g., Eur. Network Forensic Sci. Inst., Guideline for DNA Contamination Minimization in DNA Laboratories, May 10, 2023, <https://perma.cc/EQB4-55UB>. As have researchers and organizations concerned with PCR-based testing in clinical and research laboratories. E.g., World Health Organization, Dos and Don'ts for Molecular Testing (Jan. 31, 2018), <https://perma.cc/LK5B-NBC3>.

35. In the context of STR profiling (discussed in the next section), a “null allele is any allele at a microsatellite [STR] locus that consistently fails to amplify to detected levels via the polymerase chain reaction.” Elizabeth E. Dakin & John C. Avise, *Microsatellite Null Alleles in Parentage Analysis*, 93 *Heredity* 504, 504 (2004), <https://doi.org/10.1038/sj.hdy.6800545>. Such a null allele will not lead to a false exclusion if the two DNA samples from the same individual are amplified with the same primer system, but it could lead to an exclusion at one locus when searching a database of STR profiles if the database profile was determined with a different PCR kit than the one used for the crime-scene DNA.

36. Alleles at VNTR loci generally are too long to be measured precisely by electrophoretic methods—alleles differing in size by only a few repeat units may not be distinguished. Although this makes for complications in deciding whether two length measurements that are close together result from the same allele, these loci are quite powerful for the genetic differentiation of individuals, because they tend to have many alleles that occur relatively rarely in the population. At a locus with only 20 such alleles (and most loci typically have many more), there are 210 possible genotypes. With 5 such loci, the number of possible genotypes is 210^5 , which is more than 400 billion.

37. The measurement error inherent in the form of electrophoresis used is not a fundamental obstacle, but it complicates the determination of which profiles match and how often other profiles in the population would be declared to match. For a case reversing a conviction as a result of an

(STRs). STRs have very short core repeats, two to seven base pairs in length, and they typically extend for only some 50 to 350 base pairs.³⁸ Like the larger VNTRs, which extend for thousands of base pairs, STR sequences do not code for proteins or RNAs, and the ones routinely used in identity testing are four-nucleotide repeats thought to have little or no clinical value in ascertaining an individual's disease status, propensity, or behavioral characteristics.³⁹ The STR loci used in forensic identification have names such as TPOX⁴⁰ and D16S539. Figure 2 illustrates the nature of allelic variation at the latter locus, on chromosome 16.⁴¹ There are other forensic STR loci with more complicated internal patterns.⁴² Although there are fewer alleles per locus for STRs than for VNTRs, many more STRs are analyzed simultaneously (see section titled “Capillary electrophoresis and fluorescence” below).

Figure 2. Three alleles of the D16S539 STR.

Nine-repeat allele:

GATAGATAGATAGATAGATAGATAGATAGATAGATA

Ten-repeat allele:

GATAGATAGATAGATAGATAGATAGATAGATAGATAGATA

Eleven-repeat allele:

GATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATA

Note: The core sequence is GATA. The first allele listed has nine tandem repeats, the second has ten, and the third has eleven. The locus has other alleles (different numbers of repeats), shown in Figure 4.

expert's confusion on this score, see *People v. Venegas*, 954 P.2d 525 (Cal. 1998). More suitable procedures for match windows and probabilities are described in NRC II, *supra* note 7.

38. The numbers, and the distinction between “minisatellites” (VNTRs) and “microsatellites” (STRs), are not precise, but the mechanisms that give rise to the shorter tandem repeats differ from those that produce the longer ones. See Jocelyn E. Krebs et al., *Lewin's Genes XII* (2018).

39. See Amici Brief, *supra* note 19; Sara H. Katsanis & Jennifer K. Wagner, *Characterization of the Standard and Recommended CODIS Markers*, 58 J. Forensic Sci. S169 (2013), <https://doi.org/10.1111/j.1556-4029.2012.02253.x>; David H. Kaye, *Please, Let's Bury the Junk: The CODIS Loci and the Revelation of Private Information*, 102 Nw. U. L. Rev. Colloquy 70 (2007), available at <https://perma.cc/8QA3-Y3ES>. But see Mayra M. Bañuelos et al., *Associations Between Forensic Loci and Expression Levels of Neighboring Genes May Compromise Medical Privacy*, 119 Proc. Nat'l Acad. Sci. e2121024119 (2022), <https://doi.org/10.1073/pnas.2121024119>; Nicole Wyner et al., *Forensic Autosomal Short Tandem Repeats and Their Potential Association with Phenotype*, 11 Frontiers Genetics 884 (2020), <https://doi.org/10.3389/fgene.2020.00884> (concluding that no “forensic STRs . . . were found to be independently causative or predictive of disease,” but proposing that “there remains a strong chance that this inference may change in the near future”).

40. This STR is found in an intron of the thyroid peroxidase gene.

41. Names of the STRs not found in genes are in the form D#S#. The first number indicates the chromosome; the second, a location on the chromosome.

42. For examples, see Peter Gill et al., *Forensic Practitioner's Guide to the Interpretation of Complex DNA Profiles* 2–3 (2020).

Medical and human geneticists were interested in VNTRs and STRs as markers in family studies to locate the genes that are associated with inherited diseases. Papers on the potential for identity testing appeared in the early 1990s. Developmental research to pick suitable loci moved into high gear in England and other parts of Europe. Britain's Forensic Science Service applied a four-locus testing system in 1994. Then it introduced the second-generation multiplex (SGM)—for simultaneously typing six loci in 1996. These soon would be used to build the U.K.'s National DNA Database. The database system allows a computer to check the STR types of millions of known or suspected criminals against thousands of crime-scene samples. A six-locus STR profile can be represented as a string of 12 digits; each digit indicates the number of repeat units in the alleles at each locus. These discrete, numerical DNA profiles are far easier to compare mechanically than the complex patterns of fingerprints. In the United States, the FBI settled on 13 “core loci” to use in the U.S. national DNA database system.⁴³ This number is capable of distinguishing among almost everyone in the population.⁴⁴ These are often called the CODIS core loci, and an additional seven STR loci were added to the system in 2017.⁴⁵ The remainder of this section describes how laboratories typically determine which STR alleles are present after DNA from the sample is isolated.

Capillary electrophoresis and fluorescence

Determining which STR alleles are present involves ascertaining the size of the fragments that have been amplified at each STR locus. Separation according to fragment length is done in automated “genetic analyzer” machinery—a byproduct of the technology developed for the Human Genome Project that first sequenced most of the entire genome. In these machines, a long, narrow tube is filled with an entangled polymer or comparable sieving medium, and an electric

43. The federally managed Combined DNA Index System (CODIS) and related issues are discussed in the section titled “Offender and Suspect Database Searches” below; *see also* John M. Butler, *Advanced Topics in Forensic DNA Typing: Methodology* 21370 (2012).

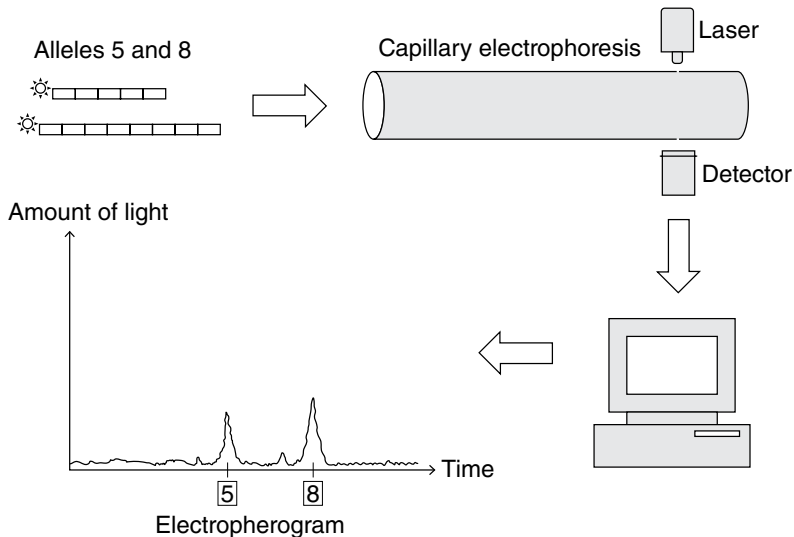
44. Usually, there are between 7 and 15 STR alleles per locus. Thirteen loci that have 10 STR alleles each can give rise to 55^{13} , or 42 billion trillion, possible genotypes—far, far greater than the U.S. population of 330 million or so. The enormous number of possible 13-locus profiles suggests that it is improbable—but not necessarily impossible—for two unrelated individuals in the country to have identical ones. Even first-degree relatives are unlikely to have the same alleles at all 13 loci. *See* section titled “Inference, Statistics, and Population Genetics in Human Nuclear DNA Testing” below. But identical twins are expected to share the same STR profile. *See* section titled “Twins” below.

45. Douglas R. Hares, *Selection and Implementation of Expanded CODIS Core Loci in the United States*, 17 *Forensic Sci. Int'l: Genetics* 33 (2015), <https://doi.org/10.1016/j.fsigen.2015.03.006>. CODIS stands for “combined DNA index system” (discussed in section titled “Offender and Suspect Database Searches” below).

field is applied to pull DNA fragments placed at one end of the tube through the medium. As with gel electrophoresis of RFLPs (see section titled “VNTRs and RFLP Testing” above), shorter fragments slip through the medium more quickly than larger, bulkier ones.

Detecting the fragments as they pass through the capillary tubes is made possible by attaching a fluorescent dye molecule to the PCR primer. As the amplified fragments move through the medium, a laser beam is sent through a small glass window in the tube, causing the dye to glow at a characteristic wavelength. The intensity of the fluorescence is recorded by a kind of electronic camera and transformed into a graph (an electropherogram), which shows a peak as the fragments with the fluorescing dye flash by. Copies of a shorter allele will pass by the window and fluoresce first; copies of a longer fragment will come by later, giving rise to another peak on the graph. Figure 3 is a sketch of how the alleles with five and eight repeats of the GATA sequence at the D16S539 STR locus might appear in an electropherogram.

Figure 3. Sketch of an electropherogram for two D16S539 alleles.



Note: One allele has five repeats of the sequence GATA; the other has eight. Each GATA repeat is depicted as a small rectangle. Although only one copy of each allele (with a fluorescent molecule, or “tag” attached) is shown here, PCR generates a great many copies from the DNA sample with these alleles at the D16S539 locus. These copies are drawn through the capillary tube, and the tags glow as the STR fragments move through the laser beam. An electronic camera measures the colored light from the tags. Finally, a computer processes the signal from the camera to produce the electropherogram.

Source: David H. Kaye, *The Double Helix and the Law of Evidence* 189, fig. 9.1 (2010).

Modern genetic analyzers produce electropherograms for many loci at once.⁴⁶ This “multiplexing” is accomplished by using dyes that fluoresce at distinct colors to label the alleles from different groups of loci. A separate set of fragments of known sizes that comigrate through the capillary function as a kind of ruler (an internal-lane size standard) to determine the lengths of the allelic fragments. Software processes the raw data to generate an electropherogram of the separate allele peaks of each color. By comparing the positions of the allele peaks to the size standard, the program determines the number of repeats in each allele. The plotted heights of the peaks (measured in relative fluorescent units, or RFUs) are proportional to the amount of the PCR product.

Complications can arise, especially with minute samples and mixtures. During PCR, the DNA strand being copied and the new strand can slip, causing “stutter” peaks that typically are one repeat away from the originating STR alleles. Fragments of DNA can fall into PCR tubes or be carried by dust particles, causing “drop in” alleles; color dyes do not fluoresce solely at the desired color and can give a signal in the adjacent color channel—a spectral artifact known as pull-up; a variety of factors can lead to unequal peaks for the pair of alleles at a locus (peak imbalance) and to missing peaks (allele dropout or even whole locus dropout).⁴⁷

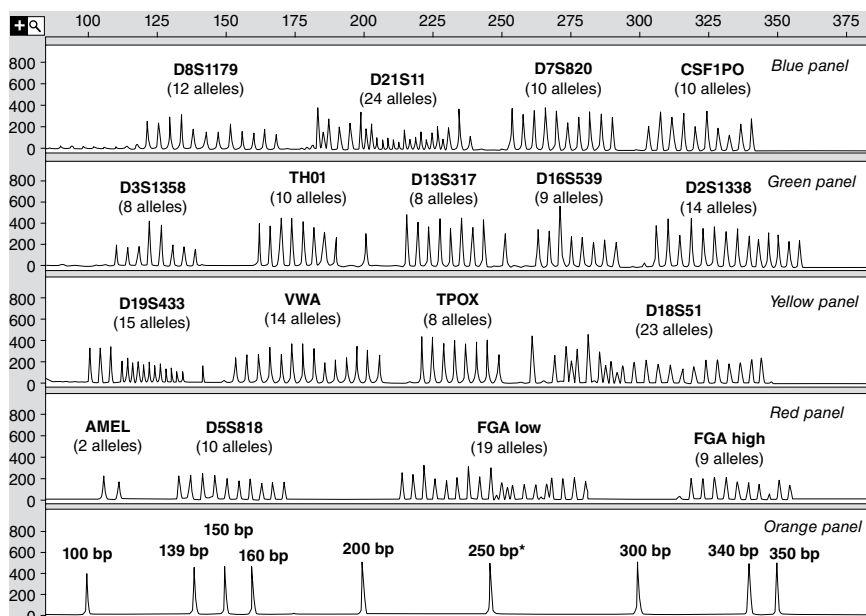
Figure 4 is an electropherogram of all 203 major alleles at 15 STR loci typed in a single multiplex PCR reaction. In addition, it shows the two alleles of the gene used to determine the biological sex of the contributor of a DNA sample.⁴⁸ An electropherogram from an individual’s DNA would have only one or two peaks at each of these 15 STR loci (depending on whether the person is homozygous or heterozygous). These “allelic ladders” aid in deciding which allele a peak from an unknown sample represents.

46. Kathryn Oostdik et al., *Developmental Validation of the PowerPlex® Fusion System for Analysis of Casework and Reference Samples: A 24-locus Multiplex for New Database Standards*, 12 *Forensic Sci. Int’l Genetics* 69 (2014), <https://doi.org/10.1016/j.fsigen.2014.04.013>; Suhua Zhang et al., *Development and Validation of a New STR 25-plex Typing System*, 17 *Forensic Sci. Int’l Genetics* 61 (2015), <https://doi.org/10.1016/j.fsigen.2015.03.008>.

47. See also sections titled “Did the Sample Contain Enough DNA?” and “Mixtures” below.

48. The amelogenin gene, which is found on the X and the Y chromosome, codes for a protein that is a major component of the tooth enamel matrix. The copy on the Y chromosome is 112 base pairs (bp) long. The copy on the X chromosome has a string of six base pairs deleted, making it slightly shorter (106 bp). A female (XX) will have one peak at 112 bp. A male (XY) will have two peaks (at 106 and 112 bp). In some populations, however, mutations that interfere with the detection of the Y chromosome by this method have been observed. These could lead to male DNA being misidentified as female on the basis of this locus alone. See Vijendra Kumar Kashyap et al., *Deletions in the Y-derived Amelogenin Gene Fragment in the Indian Population*, 7 *BMC Med. Genetics* 37 (2006), <https://doi.org/10.1186/1471-2350-7-37>.

Figure 4. Alleles of 15 STR loci and the amelogenin sex-typing test from the AmpFISTR Identifiler kit.



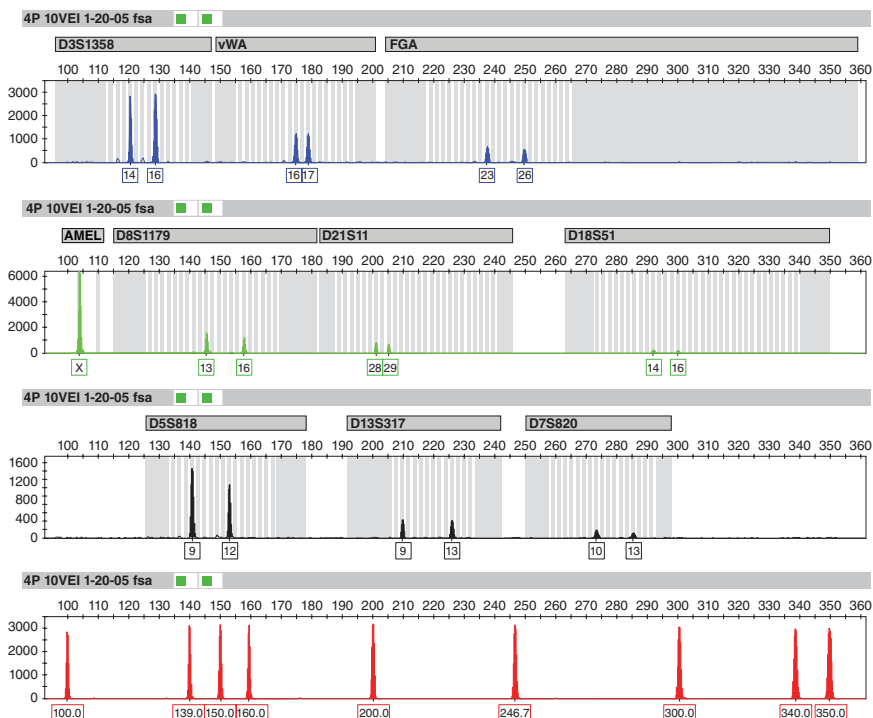
Note: The bottom panel is a sizing standard—a set of peaks from DNA sequences of known lengths (in base pairs). The numbers in the vertical axis in each panel are relative fluorescence units (RFUs) that indicate the amount of light emitted after the laser beam strikes the fluorescent tag on an STR fragment. Applied Biosystems makes the kit that produced these allelic ladders.

Source: John M. Butler, *Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers* 128 (2d ed. 2005). Copyright Elsevier 2005, with the permission of Elsevier Academic Press. John Butler supplied the illustration.

Figure 5 is an electropherogram from the vaginal epithelial cells of the body of a girl who had been sexually assaulted and killed in California.⁴⁹ It was produced for the retrial in 2008 of the defendant, who was linked to the victim by VNTR typing at his first trial in 1990.

49. *People v. Pizarro*, 12 Cal. Rptr. 2d 436 (Ct. App. 1992), *after remand*, 3 Cal. Rptr. 3d 21 (Ct. App. 2003), *review denied* (Oct. 15, 2003).

Figure 5. Electropherogram for nine STR loci of the victim's DNA in *People v. Pizarro*.



Note: The amelogenin locus and a sizing standard at the bottom also are included. Some STR loci have small peaks, indicating that there was not much PCR product for those loci, likely because of DNA degradation. All of the STR loci have two peaks, as would be expected when the source is heterozygous at those loci.

Source: Steven Myers and Jeanette Wallin, California Department of Justice, provided the image.

Miniaturization for rapid capillary electrophoresis

Miniaturized capillary electrophoresis (CE) devices have been developed for rapid detection of STRs and other genetic analyses. The mini-CE systems consist of microchannels etched on glass, silicon, or plastic wafers (“chips”) using technology borrowed from the computer industry. The microchannels are roughly the diameter of a hair. The principles of electrophoretic separation are the same as with conventional CE systems, but with microfluidic technologies, it is possible to integrate DNA extraction and PCR amplification processes with the CE separation in a single device, a so-called lab on a chip instrument that includes a cartridge with miniature pumps, mixers, valves, fluid channels

and reagent storage chambers.⁵⁰ Once a sample is added to the cartridge, all the analytical steps are performed without further human contact. These devices combine simplified sample handling with rapid analysis.

“Rapid DNA” devices have been commercialized for point-of-care medical diagnostics⁵¹ and for military and police applications. Validation for police work can be achieved by having police obtain profiles under realistic conditions and then re-running the samples with conventional equipment at established laboratories; or samples from known sources of DNA can be created for testing under experimental conditions.⁵² Several different rapid DNA machines for forensic STR profiling have been used by police departments or other agencies across the world in various ways: (1) to acquire a DNA profile during an initial encounter with a suspect; (2) to obtain the profile quickly after arrest as part of the booking process; (3) to extract and analyze DNA from crime-scene and sexual-assault-kit samples; and (4) to identify victims of mass disasters.⁵³ In the first situation, police may lack probable cause to arrest an individual but may try to conform to the requirements of the Fourth Amendment by securing valid consent to the DNA profiling.⁵⁴ Although some local police agencies—and a large prosecutors’ office—have pursued this

50. Lab-on-a-chip instruments (often called LOCs) also can use other methods than CE for analyzing DNA, proteins, or other molecules. See, e.g., Prapti Pattanayak et al., *Microfluidic Chips: Recent Advances, Critical Strategies in Design, Applications and Future Perspectives*, 25 *Microfluidics & Nanofluidics* 99 (2021), <https://doi.org/10.1007/s10404-021-02502-2>.

51. Curtis D. Chin et al., *Commercialization of Microfluidic Point-of-Care Diagnostic Devices*, 12 *Lab on a Chip* 2118 (2012), <https://doi.org/10.1039/c2lc21204h>; P. Yager et al., *Microfluidic Diagnostic Technologies for Global Public Health*, 442 *Nature* 412 (2006), <https://doi.org/10.1038/nature05064>. However, those applications would not use human STRs; instead, they would detect the DNA or RNA of pathogens.

52. E.g., Rosemary S. Turingan et al., *Developmental Validation of the ANDE 6C System for Rapid DNA Analysis of Forensic Casework and DVI Samples*, 65 *J. Forensic Sci.* 1056 (2020), <https://doi.org/10.1111/1556-4029.14286>. More detailed guidelines for validation can be found in the United Kingdom Forensic Science Regulator’s report, *Forensic Science Regulator Guidance: Methods Employing Rapid DNA Devices* § 8 (2021), <https://perma.cc/746P-9XW3> (FSR-G-229).

53. The current systems take about 90 minutes to analyze a sample.

54. Some localities go further and keep the STR data in a computer-searchable, local DNA database that they can operate independently of the databases that are part of the system of shared profile data (known as CODIS), which are administered by the FBI pursuant to the DNA Identification Act of 1994, 34 U.S.C. §§ 40702–40703 (later amended in various respects) and state statutes. Whether the indefinite retention and trawling of the local database of consent profiles for matches to future crime-scene samples exceeds the scope of the consent in a given case could be a significant question. Moreover, the non-CODIS local databases (most of which were created without rapid DNA instruments) have been criticized on other grounds. See, e.g., Jason Kreag, *Going Local: The Fragmentation of Genetic Surveillance*, 95 *B.U. L. Rev.* 1491 (2015); N.Y. City Bar Ass’n *Crim. Courts Comm.*, *Crim. Just. Operations Comm.*, and *Mass Incarceration Task Force*, *Curb-ing Unregulated Local DNA Indexing*, Apr. 16, 2021 (report supporting “an Act . . . requiring municipalities to expunge any DNA record stored in a municipal DNA identification index,” accessible via <https://perma.cc/72FC-PGL7>). In Arizona, the state Department of Public Safety chose to establish a non-CODIS database, relying on rapid DNA instruments and DNA database

practice,⁵⁵ rapid DNA technology has been more widely adopted as an adjunct to arrests. Pursuant to the Rapid DNA Act of 2017,⁵⁶ the FBI issued standards and procedures for the use of rapid DNA instruments in the booking station that allow the resulting DNA profiles to be uploaded to the National DNA Index System.⁵⁷ It is working cautiously toward the third approach, of rapid DNA analysis of crime-scene and sexual-assault-victim DNA samples.⁵⁸

Rather than face objections to rapid DNA based on Rule 702, prosecutors might choose to have accredited laboratories redo the analyses that police perform and introduce only the latter evidence. This will be possible for fresh DNA samples from defendants, but not when the crime-scene DNA has been totally consumed during the rapid testing. As with the widespread use of Breathalyzers for blood-alcohol testing,⁵⁹ ultimately courts will confront Rule 702 issues of scientific validity and proper use of the equipment. A number of organizations and individuals have expressed reservations about the current state of the technology,⁶⁰ and research is ongoing.⁶¹

management software purchased from vendors, for the benefit of localities that otherwise might have been tempted to establish such systems on their own. Kreaq, *supra*, at 1517–19.

55. In Orange County, California, prosecutors with rapid DNA machines in their basement assembled a significant database by “offer[ing] defendants accused of misdemeanors and infractions a deal: give the prosecutor’s office your DNA, and the office will offer you leniency in your criminal case.” Andrea Roth, “*Spit and Acquit*”: Prosecutors as Surveillance Entrepreneurs, 107 Cal. L. Rev. 405, 408 (2019). California Senate Bill 1228 (adopted 2022) curbs such practices.

56. Pub. L. No. 115–50, 131 Stat. 1001 (codified at 34 U.S.C. §§ 12591, 12592, 40702, 40703).

57. See FBI, Guide to All Things Rapid DNA, Jan. 27, 2022, <https://le.fbi.gov/file-repository/rapid-dna-guide-january-2022.pdf/view>. Previously, only qualified state laboratories could submit profiles.

58. *Id.*

59. See 1 McCormick on Evidence, *supra* note 1, § 205.1, at 1309–10 n.12.

60. There are concerns over fully automated analysis of small samples and mixtures because of sample consumption, reduced sensitivity relative to full-scale equipment, and the integrated software for interpretation. Discussions of the state of the art and research and administrative recommendations can be found in, for example, Erik Dalin et al., Biology Section, Nat’l Forensic Centre, Swedish Police Authority, Rapid DNA: A Summary of Available Rapid DNA Systems (2022), <https://perma.cc/87XS-WQ9W>; Douglas R. Hares et al., *Rapid DNA for Crime Scene Use: Enhancements and Data Needed to Consider Use on Forensic Evidence for State and National DNA Databasing—An Agreed Position Statement by ENFSI, SWGDAM and the Rapid DNA Crime Scene Technology Advancement Task Group*, 48 Forensic Sci. Int’l: Genetics 102349 (2020), <https://doi.org/10.1016/j.fsigen.2020.102349>; Non-CODIS Rapid DNA Best Practices/Outreach and Courtroom Considerations Task Group, Texas Forensic Sci. Comm’n, Non-CODIS Rapid DNA Considerations and Best Practices for Law Enforcement Use, Sept. 16, 2019, <https://perma.cc/X3VV-RCUB>.

61. E.g., Belinda Martin et al., *Analysis of Rapid HIT Application to Touch DNA Samples*, 67 J. Forensic Sci. 1233 (2022), <https://doi.org/10.1111/1556-4029.14964> (“not fit for the routine analysis of touch DNA samples in forensic casework”); Denise Ward et al., *Analysis of Mixed DNA Profiles from the RapidHIT™ ID Platform Using Probabilistic Genotyping Software STRmix™*, 58 Forensic Sci. Int’l: Genetics 102664 (2022), <https://doi.org/10.1016/j.fsigen.2022.102664> (“good discrimination power [when STR data was analyzed with external software] but less than those produced via the standard laboratory workflow,” as would be “expected . . . for the advantages of speed and portability”).

Sequence-Specific Probes and Microarrays

Simple sequence variation, such as that for the GC locus (see section titled “Chromosomes” above), is conveniently detected using sequence-specific probes (defined in the sections on RFLP-VNTR and STR profiling). With GC typing, for example, probes for the three common alleles are attached at three separate spots on a membrane. Copies of the variable sequence region of the GC gene in the crime-scene sample are made with PCR. These copies (in the form of single strands) are poured onto the membrane. Whichever allele is present in a copy will cause it to stick to a corresponding, immobilized probe strand. A chemical “label” that catalyzes a color change at the spot where the trace-evidence allele binds to its probe can be attached when the copies are made. A colored spot showing that the allele is present thus should appear on the membrane at the location of the probes that correspond to this particular allele. If only one allele is present in the crime-scene DNA (because of homozygosity), there will be no color change at the spots where the other probes are located. If two alleles are present (heterozygosity), the corresponding two spots will change color.

This method was used before STR profiling, with only a few loci.⁶² In that form, it lacked the discriminating power of STRs and VNTRs, but, being PCR-based, it was more sensitive, and easier and quicker to perform than RFLP-VNTR profiling. Admissibility followed.⁶³

The approach can be miniaturized and automated by embedding probes for many loci on a chip.⁶⁴ Such a “microarray” for DNA-sequence detection usually consists of a two-dimensional grid of many thousands of microscopic spots on a silicon, glass, plastic, or paper surface. Each spot contains many copies of a probe with its own particular sequence, tethered to the surface at one end. A solution containing copies of single-stranded target DNA fragments⁶⁵ (with attached fluorescent tags or dyes) is washed over the microarray surface. As usual, the probes on the array bind to copies with the complementary sequence. The spots

62. Indeed, in what probably was the first criminal case in the United States in which DNA evidence was admitted, a single-locus test that suggested that the owners of a nursing home, although found guilty of negligent homicide for the starvation of an elderly patient, did not swap the organs of the victim with those of another cadaver in an alleged attempt to cover up the cause of death. Stephen Michaud, *DNA Detectives*, N.Y. Times Mag., Nov. 6, 1988, § 6, at 70, available at <https://www.nytimes.com/1988/11/06/magazine/dna-detectives.html>.

63. On the development and judicial reception of the “dot blot” or “reverse dot blot” tests and the “polymarker” system, see Kaye, *supra* note 1, at 180–87.

64. See, e.g., Johannes Wöhrle et al., *Digital DNA Microarray Generation on Glass Substrates*, 10 Sci. Reps. 5770 (2020), <https://doi.org/10.1038/s41598-020-62404-1>.

65. Amplification (including isothermal methods that do not require the heating and cooling cycles of conventional PCR) of the target DNA can be built into the microarray. Roger Bumgarner, *Overview of DNA Microarrays: Types, Applications, and Their Future*, 101 Current Protocols in Molecular Biology 22–1 (2013), <https://doi.org/10.1002/0471142727.mb2201s101>.

that capture the targeted DNA are identified, indicating the presence of those sequences in the sample.

DNA microarrays have broad applications in biology and medicine.⁶⁶ Forensic applications include sequencing human mitochondrial DNA (see section titled “Mitochondrial DNA” below) and resolving individual genotypes within complex DNA mixtures,⁶⁷ but the most well known application comes from SNP chips with enough different probes to detect half-a-million or more different known SNPs distributed throughout the human genome.⁶⁸ After biotechnology companies produced these microarrays for genomics research and health applications,⁶⁹ direct-to-consumer genetic testing companies popularized them for ancestry and other “recreational genetics” activities. As more and more people interested in locating possibly unknown relatives elected to share certain genomic information in a public database, trawling the database for putative relatives of perpetrators of cold cases to infer the identity of the unknown source of the crime-scene DNA proved spectacularly successful. This procedure of investigative genetic genealogy is the subject of a later section by that name.

Sequencing and SNPs

As shown in Figure 1 (see section titled “The DNA Molecule” above), DNA contains a sequence of paired letters—the nucleotides A, T, G, and C—laid out

66. They permit large-scale population studies—for example, to determine how often individuals with a particular mutation develop breast cancer, or to identify the changes in gene sequences that are most often associated with particular diseases. Microarrays also are used in studies of variation in the number of copies of certain genes in different people’s genomes (copy number variation). They are used to study the extent to which certain genes are turned on or off in cells and tissues. For this purpose, instead of isolating DNA from the samples, a transcript of the DNA (mRNA, *supra* section titled “Chromosomes”) is isolated and measured. They provide clinical diagnostic tests for diseases. “Biosensor” microarrays detect pathogens and other targets.

67. Lev Voskoboinik et al., *SNP-Microarrays Can Accurately Identify the Presence of an Individual in Complex Forensic DNA Mixtures*, 16 *Forensic Sci. Int’l: Genetics* 208 (2015), <https://doi.org/10.1016/j.fsigen.2015.01.009> (proof of concept).

68. Another proposed use is providing SNP data for better discerning the distinct contributors to DNA mixtures. Nils Homer et al., *Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays*, 4 *PLoS Genetics* e1000167 (2008), <https://doi.org/10.1371/journal.pgen.1000167>. But see Rosemary Braun et al., *Needles in the Haystack: Identifying Individuals Present in Pooled Genomic Data*, 5 *PLoS Genetics* e1000668 (2009), <https://doi.org/10.1371/journal.pgen.1000668>; Peter M. Visscher & William G. Hill, *The Limits of Individual Identification from Sample Allele Frequencies: Theory and Statistical Analysis*, 5 *PLoS Genetics* e1000628, <https://doi.org/10.1371/journal.pgen.1000628>.

69. One study of 3,000 Europeans used a commercial microarray with over half a million SNPs “to infer [the individuals’] geographic origin with surprising accuracy—often to within a few hundred kilometers.” John Novembre et al., *Genes Mirror Geography Within Europe*, 456 *Nature* 98, 98 (2008), <https://doi.org/10.1038/nature07331>.

in a spiraling kind of line. Broadly, sequencing refers to ascertaining the precise order of the DNA letters along each strand.⁷⁰ The sequence-specific probes of the previous section could be said to sequence short segments of DNA in those instances in which they bind to the complementary sequence in the target DNA. But that is not how DNA sequencing machines work. They determine the order of base pairs one at a time, somewhat like spelling out the letters of a word before one recognizes it as a word. This sequencing is the subject of this section. It can be done for short stretches of DNA or, with enough effort and ingenuity, and in a roundabout way, for the whole genome, ultimately revealing the order of the base pairs at every gene, every STR, and every other kind of locus.

Sequencing methods

Sequencing technology is hardly new. In the 1970s, biochemists developed two PCR-related methods to sequence DNA.⁷¹ For the next 30 years, researchers applied “Sanger sequencing” to decipher complete genes and, later, entire genomes of organisms. It is still used to sequence particular regions (up to a thousand or so base pairs). Basically, it uses PCR to synthesize a series of complementary DNA strands with color-coded nucleotides that indicate whether the nucleotide in the template strand is an A, T, G, or C.⁷² The multinational, multi-year Human Genome Project, completed in the early 2000s, employed factory-like systems with hundreds of sequencing machines to Sanger sequence most of the 3.2 billion base pairs in the haploid genome from several anonymous

70. Because of the strict A-T, G-C pairing rule, it is not necessary to write two letters to designate a pair; the sequence of bases on one strand implies the sequence on the other.

71. Allan M. Maxam & Walter Gilbert, *A New Method for Sequencing DNA*, 74 Proc. Nat’l. Acad. Sci. 560 (1977), <https://doi.org/10.1073/pnas.74.2.560>; Frederick Sanger et al., *DNA Sequencing with Chain-terminating Inhibitors*, 74 Proc. Nat’l. Acad. Sci. 5463 (1977), <https://doi.org/10.1073/pnas.74.12.5463>.

72. As with ordinary PCR amplification, the target DNA is unzipped and copied—but the polymerase chain reaction stops prematurely because some “chain-terminating” nucleotides (the A, T, C, and G units) are added to the PCR mixture. The chemically modified nucleotides are missing the molecular “hook” that lets the next unit attach itself to the growing chain, and they have a fluorescent dye attached to them (with a different color for A, T, G, and C units). The chain reaction continues adding nucleotides until it happens to add a modified one. This process is repeated many times to generate fragments that extend to all possible stopping points. The smallest ones will consist of only one terminating unit after the primer. The next smallest will have two nucleotides, and so on, for every base pair in the original target DNA fragment. Capillary electrophoresis (*supra* section titled “STRs and Capillary Electrophoresis”) separates the synthesized fragments from smallest to largest, with the results displayed as colored peaks in the electropherogram (also called a chromatogram). The order of the colored peaks gives the sequence of the base pairs in the target DNA. The whole process is an example of sequencing by synthesis.

individuals.⁷³ Despite the ultimate success, the sequencing efforts were technically demanding, expensive, time-consuming, and not suitable for regions on the chromosomes that are replete with repetitive DNA.⁷⁴

In 2004, the National Human Genome Research Institute announced funding for research leading to the “\$1,000 genome,” which would permit sequencing an individual’s genome for medical diagnosis and improved drug therapies. The goal was essentially met in 2014 thanks to methods known as next-generation sequencing (NGS) or massively parallel sequencing (MPS).⁷⁵ MPS methods simultaneously read millions of fragments and then use computer-intensive alignment programs to stitch the reads together to give the whole genome sequence. The genomes can be small (as with bacteria), large (as with humans), or larger still (as with some plants and other organisms). The methods also can be applied to selected parts of a genome, making them attractive for forensic and diagnostic purposes.

Commercially available sequencers implement the MPS approach in different ways.⁷⁶ One popular system is another form of sequencing by synthesis. First, to prepare the sample for sequencing, all the DNA is cut into many little pieces, or the targets are amplified as small pieces, creating a “library” of fragments. Then “adapters” are added to the ends of the pieces in this library. These adapters are like tiny handles that help immobilize the DNA to facilitate sequencing. Second, instead of running PCR in a tube, the DNA is loaded onto a glass “flow cell” that has millions of tiny wells on its surface. Each of these wells can grab an adapter to capture a single piece of DNA from the library. Third, the fragments in the wells are amplified with PCR, resulting in a cluster of identical single-stranded DNA in a given well. Fourth, chemically modified nucleotides bind to the DNA template strand. As in Sanger sequencing, each nucleotide contains a fluorescent tag, but the terminator is removed after the color at each well is detected, allowing the next base to bind. The reactions are repeated hundreds of times to trace the sequence of nucleotides in the growing chain. The cycle of one-pair-at-a-time synthesis occurs in millions of wells at once. Finally, bioinformatics software fits the jigsaw puzzle of millions of pieces together.⁷⁷

73. Int’l Human Genome Sequencing Consortium, *Finishing the Euchromatic Sequence of the Human Genome*, 431 *Nature* 931 (2004), <https://doi.org/10.1038/nature03001>; Int’l Human Genome Sequencing Consortium: *Initial Sequencing and Analysis of the Human Genome*, 409 *Nature* 860 (2001), <https://doi.org/10.1038/35057062>; cf. J. Craig Venter et al., *The Sequence of the Human Genome*, 291 *Science* 1304 (2001), <https://doi.org/10.1126/science.1058040>.

74. More efficient methods have enabled researchers to fill in the gaps, correct errors, and release “a truly complete human reference genome.” Sergey Nurk et al., *The Complete Sequence of a Human Genome*, 376 *Science* 44 (2022), <https://doi.org/10.1126/science.abj6987>.

75. Erwin L. van Dijk et al., *Ten Years of Next-Generation Sequencing Technology*, 30 *Trends in Genetics* 418 (2014), <https://doi.org/10.1016/j.tig.2014.07.001>.

76. Barton E. Slatko et al., *Overview of Next-Generation Sequencing Technologies*, 122 *Current Protocols in Molecular Biology* e59 (2018), <https://doi.org/10.1002/cpmb.59>.

77. With MPS, each base has to be read not just once, but at least several times in the overlapping segments to ensure accuracy in fitting the reads together.

The MPS sequencing-by-synthesis method described above sometimes is cited as an example of second-generation sequencing. Its reads are much shorter and more prone to error (per read) than Sanger sequencing,⁷⁸ but the speed and cost are enormous improvements for sequencing large genomes.

The sequencing part of the process (the steps starting with insertion of the library of DNA fragments into the flow cell) need not be applied to whole genomes. The library can consist of copies of short segments derived from any interesting sequences.⁷⁹ MPS thus supports targeted sequencing of many polymorphisms of forensic interest all at once.⁸⁰

A second example of an MPS technology has been called third- or even fourth-generation sequencing. It produces long reads from a single DNA molecule without having to amplify it by PCR and to attach and illuminate fluorescent labels. Proteins embedded into a biological membrane create exquisitely small holes, called nanopores. (A human hair is on the order of 100,000 nanometers wide; a DNA molecule is 2–12 nanometers wide.) Either single- or double-stranded DNA can be moved through multiple nanopores of a flow cell at a constant rate for tens of thousands of nucleotides. As a DNA molecule is ratcheted through a pore, it interrupts a flow of electrical current to a sensor chip. The disruption produces a disturbance in the current that is characteristic of the type of nucleotide (A, T, G, or C) passing through. Software decodes the disturbance to determine the DNA sequence as the nucleotides flow by.

With such long-read sequencing, assembling the overlapping reads is easier, just as putting together a jigsaw puzzle with fewer, larger pieces is easier than fitting together numerous smaller ones. Moreover, large insertions or deletions and repetitive regions can be detected more easily. Although the longer reads can have a higher error rate, ongoing technological improvements are closing the accuracy gap with the sequencing-by-synthesis versions of MPS.⁸¹

78. Sanger sequencing platforms “offer well-trusted, up to 1-kbp long reads . . . and are still considered the gold standard for sequencing. They are routinely used for clinical gene tests or validations. However, their low-throughput limits them to single or small batches of targets.” Adam Ameur et al., *Single-Molecule Sequencing: Towards Clinical Applications*, 37 *Trends in Biotech.* 72 (2019), <https://doi.org/10.1016/j.tibtech.2018.07.013>.

79. Panels of genes with variants that are known to promote the growth of cancer cells can be targeted to see which mutations are present in cancer patients. See, e.g., Masayuki Nagahashiet et al., *Next Generation Sequencing-based Gene Panel Tests for the Management of Solid Tumors*, 110 *Cancer Sci.* 6 (2019), <https://doi.org/10.1111/cas.13837>.

80. Peter de Knijff, *From Next Generation Sequencing to Now Generation Sequencing in Forensics*, 38 *Forensic Sci. Int'l: Genetics* 175 (2019), <https://doi.org/10.1016/j.fsigen.2018.10.017>.

81. E.g., Søren M. Karst et al., *High-Accuracy Long-Read Amplicon Sequences Using Unique Molecular Identifiers with Nanopore or Pacbio Sequencing*, 18 *Nature Methods* 165 (2021), <https://doi.org/10.1038/s41592-020-01041-y>.

High-throughput sequencing technologies have demonstrated their usefulness in a wide range of research,⁸² clinical,⁸³ and public health⁸⁴ applications. A famous example is the whole-genome sequencing of highly degraded ancient DNA (aDNA) of extinct species, including woolly mammoths, cave bears, Neanderthals, and Denisovans, as well as more modern human populations, from Vikings to Paleo-Inuit.⁸⁵

The range of forensic applications

Of course, different applications within the broad category of massively parallel or next-generation sequencing can use different techniques, and particular forensic applications need to be validated to demonstrate that they are fit for their intended use. Research has shown that MPS can be used for the following:

- to increase the power of STR loci to distinguish among individuals by detecting sequence differences within the length-based alleles as measured by capillary electrophoresis;⁸⁶

82. E.g., Vivian Marx, *Method of the Year: Long-read Sequencing*, 20 *Nature Methods* 6 (2023), <https://doi.org/10.1038/s41592-022-01730-w>; Katia Nones & Ann-Marie Patch, *The Impact of Next Generation Sequencing in Cancer Research*, 12 *Cancers (Basel)* 2928 (2020), <https://doi.org/10.3390/cancers12102928>.

83. Elaine Hsu et al., *Rapid Pathogen Detection by Metagenomic Next-Generation Sequencing of Infected Body Fluids*, 27 *Nature Med.* 115 (2021), <https://doi.org/10.1038/s41591-020-1105-z>; F. Mosele et al., *Recommendations for the Use of Next-Generation Sequencing (NGS) for Patients with Metastatic Cancers: A Report from the ESMO Precision Medicine Working Group*, 31 *Annals Oncology* 1491 (2020), <https://doi.org/10.1016/j.annonc.2020.07.014>. But see Francois Balloux et al., *From Theory to Practice: Translating Whole-Genome Sequencing (WGS) into the Clinic*, 26 *Trends in Microbiology* 1035 (2018), <https://doi.org/10.1016/j.tim.2018.08.004> (discussing obstacles to more routine use); Nagahashiet et al., *supra* note 79.

84. E.g., Rowena A. Bull et al., *Analytical Validity of Nanopore Sequencing for Rapid SARS-CoV-2 Genome Analysis*, 11 *Nature Commun.* 6272 (2020), <https://doi.org/10.1038/s41467-020-20075-6>; David F. Nieuwenhuijse et al., *Towards Reliable Whole Genome Sequencing for Outbreak Preparedness and Response*, 23 *BMC Genomics* 569 (2022), <https://doi.org/10.1186/s12864-022-08749-5>; Joshua Quick et al., *Real-time, Portable Genome Sequencing for Ebola Surveillance*, 530 *Nature* 228 (2016), <https://doi.org/10.1038/nature16996>. By sequencing entire bacterial genomes, researchers can rapidly differentiate organisms that have been genetically modified for biological warfare or terrorism from routine clinical and environmental strains. B. La Scola et al., *Rapid Comparative Genomic Analysis for Clinical Microbiology: The Francisella Tularensis Paradigm*, 18 *Genome Res.* 742 (2008), <https://doi.org/10.1101/gr.071266.107>.

85. Elizabeth D. Jones, *Ancient DNA: The Making of a Celebrity Science* (2022); Andrew Curry, *Ancient DNA Pioneer Svante Pääbo Wins Nobel*, 378 *Science* 12 (2022), <https://doi.org/10.1126/science.adf1845>.

86. For example, one allele might have nine copies of a GATA repeat (as in Figure 2), while another might have eight GATA repeats and one GTTA instead of a GATA. Their PCR products, being the same length, would migrate through the gel at the same speed, so they would show up as a single peak on an electropherogram. So would two alleles—from a different individual—that are

- to better resolve mixtures into the components coming from different contributors (as discussed in the section titled “Mixtures” below);
- to develop investigative leads based on biogeographic ancestry and physical characteristics such as eye color;⁸⁷
- to ascertain the tissues that the DNA in crime-scene samples came from when the whole cells are not present in the samples;⁸⁸
- to sequence mitochondrial DNA;⁸⁹ and even
- to distinguish between identical twins.⁹⁰

Moreover, as explained in the next section, the ability of MPS to find point mutations (insertions, deletions, or substitutions of single base pairs) opens up the possibility of using well-chosen SNPs as a replacement for (or at least a supplement to) STR profiling for the purpose of individual identification, that is, of determining whose DNA is in a crime-scene sample.⁹¹ Biotechnology companies have developed kits for preparing the libraries for all these purposes,⁹² and in 2024, MPS results were held to be admissible for the first time in the United States.⁹³

SNPs as identification loci

Single-nucleotide polymorphisms (SNPs) are the most abundant variations in the human genome. SNPs are shorter than STRs, making it possible to type degraded samples that regular STR profiling cannot handle because the

made up of nine GATA repeats. PCR-CE would not distinguish between these two individuals, but the more detailed sequence information could. The additional information also can improve mixture resolution (*see infra* section titled “Mixtures”).

87. *See infra* section titled “Investigative DNA Analysis.”

88. *See infra* section titled “DNA and RNA Typing of Bodily Fluids or Tissues.”

89. *See infra* section titled “Mitochondrial DNA.”

90. *See infra* section titled “Twins.”

91. *See, e.g.,* Christopher P. Phillips et al., *A Compilation of Tri-allelic SNPs from 1000 Genomes and Use of the Most Polymorphic Loci for a Large-Scale Human Identification Panel*, 46 *Forensic Sci. Int'l: Genetics* 102232 (2020), <https://doi.org/10.1016/j.fsigen.2020.102232>.

92. Penelope R. Hadrill. *Developments in Forensic DNA Analysis*, 5 *Emerging Topics Life Sci.* 381 (2021), <https://doi.org/10.1042/ETLS20200304>.

93. *People v. Chavez*, No. BF187877A (Kern Cnty. Cal. Super. Ct. Jan. 22, 2024). Defendant was convicted of first-degree murders of two homeless women. In this unreported case, “[f]orensic evidence, including DNA and fingerprints, directly linked [the defendant] to both deaths.” Office of the Dis. Atty., Press Release, Jan. 24, 2024, <https://perma.cc/4FX8-3K6G>. Admission followed an “extensive” pretrial hearing on results obtained with a “MiSeq FGx System.” *Id.* This “instrument [can] interrogate up to 96 combined SNP and STR libraries in a single run” using a massively parallel sequencing-by-synthesis procedure. Verogen MiSeq FGx Sequencing System for Next-generation Sequencing of Forensic Genomics Libraries, <https://perma.cc/LUF7-YMSP>.

fragments are shorter than the STR alleles.⁹⁴ Indeed, even with undegraded samples, STR analysts must cope with phenomena such as stutter peaks that are near a true peak but might be confused with a peak from a different allele. SNP analysis would obviate this problem with electrophoretic-STR data.

On the other hand, a SNP locus has fewer alleles than an STR locus (usually only two). Hence, it takes more of them to achieve high levels of discrimination. MPS systems solve this problem because they can analyze many separate SNP loci at once. Data on how frequently major SNP alleles occur in various population groups have been collected, so the significance of a match in a multilocus profile can be assessed numerically.⁹⁵ As such, some scientists see STR profiling as outdated—but recommend including STRs in the sequencing process if only to avoid having to redo the DNA databases of STR profiles of convicted offenders and arrestees.

Recent work with SNPs for individual identification focuses on using more than one SNP at a time for a marker. Two or three SNPs can be chosen within a short segment of DNA (smaller than 300 base pairs, for example). Such short blocks of DNA are very likely to be inherited as a package from one parent, and MPS targeted to them can identify this “microhaplotype” in a single sequence run. A pair of SNPs has more possible types than a single SNP, so each microhaplotype is more informative than a single-SNP locus. Panels of microhaplotypes have been studied for individual identification and other forensic-science applications.⁹⁶ With a sufficient number of SNPs, the power to discriminate between two randomly selected individuals is superior to that of conventional STR profiling.⁹⁷

94. Moreover, SNPs have advantages for forensic tests that do not make direct comparisons of samples of known (from a suspect, for example) and unknown (from a crime scene) origin. They tend to be specific to certain populations, so they can be used to infer ancestry. *See infra* section titled “Biogeographic Ancestry Testing.” Most have far lower mutation rates, which is advantageous in parentage and other kinship testing.

95. *See infra* sections titled “Could a Close Relative Be the Source?” and “Frequencies, Probabilities, and Prejudice.”

96. To reduce the possibility of conveying significant information about disease status or susceptibility, microhaplotype loci can be limited to SNPs located far from genes. Ones that are very close to certain genes would tend to be inherited along with the gene, potentially allowing them to be used as markers for disease-causing gene mutations.

97. *E.g.*, Kenneth K. Kidd et al., *Evaluating 130 Microhaplotypes Across a Global Set of 83 Populations*, 29 *Forensic Sci. Int'l: Genetics* 29 (2017), <https://doi.org/10.1016/j.fsigen.2017.03.014>; Aliye Kureshi et al., *Construction and Forensic Application of 20 Highly Polymorphic Microhaplotypes*, 7 *Royal Soc'y Open Sci.* 191937 (2020), <https://doi.org/10.1098/rsos.191937>; Jing-Bo Pang et al., *A 124-plex Microhaplotype Panel Based on Next-generation Sequencing Developed for Forensic Applications*, 10 *Sci. Reps.* 1945 (2020), <https://doi.org/10.1038/s41598-020-58980-x>; Andrew J. Pakstis et al., *The Population Genetics Characteristics of a 90 Locus Panel of Microhaplotypes*, 140 *Hum. Genetics* 1753 (2021), <https://doi.org/10.1007/s00439-021-02382-0>.

Summary

DNA contains the genetic information of an organism. In humans, most of the DNA is found in the cell nucleus, where it is organized into separate chromosomes. Each chromosome is like a book, and each cell has the same set of books of various sizes. There are two copies of each book (a “diploid” genome)—one copy from the father, one from the mother (the two “haploid” genomes). Thus, there are two copies of the book entitled “Chromosome One,” two copies of “Chromosome Two,” and so on. The text is mostly the same in a pair of books, but sexual reproduction results in important differences as well as inconsequential ones.

Genes are like paragraphs in the books. They encode different proteins and RNAs. Parts of the DNA text appear to have no coherent message, but some of the noncoding DNA affects the production of the gene products. Two individuals sometimes have different versions (alleles) of the same gene as well as variants of the text outside of (or interrupting) the genes. Some alleles result from the substitution of one letter for another. These are SNPs. Others come about from the insertion or deletion of single letters, and still others represent a kind of stuttering repetition of a string of extra letters. These are the VNTRs and STRs.⁹⁸ The locations within a chromosome where these interpersonal variations occur are called loci.

Historically, forensic DNA typing relied primarily on the length differences of a small number of VNTR and then a larger number of STR loci.⁹⁹ STR profiling depends on a chemical reaction for generating a great many copies of DNA segments (PCR amplification), followed by separation of the copies according to their length (capillary electrophoresis). We can refer to this well-established technology as PCR-CE for STRs, or even more compactly, as STR-CE profiling. Modern sequencing technology enables more and improved SNP-based loci for identity determinations and other forensic-science applications.

What Establishes That a Genetic System Is Valid for Identification?

Regardless of the kind of genetic system used for typing—STR, SNP, or still other polymorphisms—some general principles and questions can be applied to judge its validity.¹⁰⁰ First, the nature of the polymorphism should be well

98. In addition to the 23 pairs of “books” (chromosomes) in the cell nucleus, other scraps of DNA “text” reside in each of the mitochondria, the power plants of the cell. See section titled “Mitochondrial DNA” below.

99. For a short time, a few SNPs in protein-coding DNA also were used.

100. In addition to the suggestions on studying validity in this section, see section titled “Validation” below.

characterized. Is it a sequence polymorphism or a length polymorphism? Where is it situated within the genome? This information should be in the published literature or in archival genome databanks.

Second, the published scientific literature can be consulted to verify claims that a particular method of analysis can produce accurate profiles under various conditions. Ideally, these claims will rest on a detailed understanding of how the analytical procedure works—that is, of how measurements are made and how inferences are drawn from these measurements—as well as empirical studies of the results achieved with the methods under experimental conditions or in casework.¹⁰¹ Although such validation studies have been conducted for all the systems described here, determining the point at which the validation of a particular system is sufficiently extensive and convincing to pass scientific muster may well require expert assistance.¹⁰² The standards for validation that have been issued by private standards-developing organizations or expert groups often have only general requirements.¹⁰³

Finally, the population genetics of the system should be characterized. As new systems are discovered, researchers typically analyze convenient collections of DNA samples from various human populations and publish studies of the relative frequencies of each allele in these population samples.¹⁰⁴ These studies give a measure of the extent of variability at the polymorphic locus in the various populations, and thus of the potential probative power of the marker for distinguishing among individuals.

At this point, the capability of the widely used PCR-CE procedures to ascertain STR genotypes accurately cannot be doubted.¹⁰⁵ But the technology has its limits, and a laboratory should be able to show that it is operating within the limits

101. *See id.*

102. *See* section titled “Relevant Expertise” above.

103. When it comes to specifying what level of performance an analytical method must be shown to possess, many standards are vague or vacuous. For instance, the FBI’s Scientific Working Group on DNA Analysis Methods conflates accuracy with repeatability and reproducibility of measurements and cursorily asserts that these properties “should be evaluated.” SWGDAM, Validation Guidelines for DNA Analysis Methods §§ 3.5.1 & 3.5.2 (Dec. 5, 2016). ANSI/ASB Standard 038 (2020), entitled “Standard for Internal Validation of Forensic DNA Analysis Methods,” likewise contains no minimum levels for accuracy and reproducibility, and its “requirements” barely go beyond the command that “[t]he laboratory shall conduct internal validation studies on all forensic DNA analysis methodologies prior to implementation.” More detailed standards for particular techniques are in progress.

104. On the populations that should be sampled, *see* section titled “Could an Unrelated Person Be the Source?” below.

105. *See, e.g.,* President’s Council of Advisors on Sci. & Tech. (PCAST), Report to the President: Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods, Sept. 2016, at 75, <https://perma.cc/R76Y-7VU> (“single-source samples or simple mixtures of two individuals, such as from many rape kits, is an objective method that has been established to be foundationally valid”); *accord*, John M. Butler et al., DNA Mixture Interpretation: A NIST Scientific Foundation Review, at 3 & 7, June 2021, NISTIR 8351-DRAFT; *see also supra* Introduction.

for which it has been tested (or if not, to explain why the change in circumstances does not matter). If small samples are at issue, it will be important to ask what validity studies show about the performance of the system as sample size is reduced. If samples include DNA mixtures, the question of whether the validity studies show that the method works well enough with similarly complex samples will arise.¹⁰⁶

As next-generation technologies are introduced in court, the same basic questions will need to be answered: What is the principle of the new technology? Is it simply an extension of existing technologies, or does it invoke entirely new concepts? Is the new technology used in research or clinical applications independent of forensic science? If so, are there grounds for concern that the new technology has limitations that might affect its application in the forensic sphere? Finally, what testing has been done to establish that the new technology is valid and reliable when used on forensic samples? For massively parallel sequencing and microarray technologies, the questions may be directed as well to the bioinformatics methods used to analyze and interpret the raw data. Obtaining answers to these questions will likely require input both from experts involved in technology development and application and from knowledgeable forensic experts.

Of course, the fact that scientists have shown that it is possible to extract DNA and to analyze it in a way that bears on the issue of identity does not mean that a particular laboratory has adopted a suitable protocol, is generally proficient in following it, and has implemented the procedure correctly in the case at bar. These more case-specific issues are considered next.

Sample Collection and Laboratory Performance

Sample Collection and Contamination

The primary determinants of whether DNA typing can be done on any particular sample are (1) the quantity of DNA present in the sample and (2) the extent to which it is degraded. Generally speaking, if a sufficient quantity of reasonable-quality DNA can be extracted from a crime-scene sample, no matter what the nature of the sample, DNA typing can be done. Thus, DNA typing has been performed on old blood stains, semen stains, vaginal swabs, hair, bone, bite marks, cigarette butts, drinking cups, guns, garments, urine, and fecal material. This section discusses what constitutes sufficient quantity and reasonable quality in the context of capillary electrophoresis of PCR-amplified short tandem repeats (STR-CE profiling), which has been the predominant method for forensic DNA identification for more than thirty years. Complications from

106. See section titled “Validation of PGS Programs” below.

contaminants and inhibitors also are discussed, and emerging alternatives to STR-CE profiling are noted. The treatment of samples that contain DNA from two or more contributors is discussed in the section titled “Mixtures” below.

Did the Sample Contain Enough DNA?

Amounts of DNA present in some typical kinds of samples vary from a trillionth or so of a gram (a picogram) for a hair shaft to several millionths of a gram (micrograms) for a postcoital vaginal swab.¹⁰⁷ Most PCR test protocols recommend samples on the order of 1 billionth of a gram (1 nanogram) for optimal yields. Normally, the number of amplification cycles for nuclear DNA is limited to 28, 29, or 30 to avoid detecting alleles from less than about 10 to 15 cell equivalents of DNA (66 to 100 picograms).¹⁰⁸

Procedures for STR-CE typing of still smaller samples—down to a single cell’s worth of nuclear DNA—have been studied. These have been shown to work, to some extent, with trace or contact DNA left on the surface of an object such as the steering wheel of a car.¹⁰⁹ Improved reaction chemistries, amplification to a greater number of PCR cycles,¹¹⁰ and detection systems that provide greater sensitivity have made detection of PCR products associated with single

107. A combination of chemical and physical methods permits DNA to be isolated from the other chemicals in a specimen collected from a crime scene or an individual. The basics are fairly standard in academic research (*see generally* Akash Gautam, DNA and RNA Isolation Techniques for Non-Experts (2022)), but scientists in a forensic DNA laboratory often face a high volume of samples, a variety of sample types, limited quantities of biological material, DNA degradation, chemical inhibitors, and environmental contaminants that can interfere with later analysis. Hence, there is no single best DNA extraction method. The method of choice often depends on the biological sample concerned. Kevin Wai Yin Chong et al., *Recent Trends and Developments in Forensic DNA Extraction*, 3 WIREs Forensic Sci. e1395 (2020), <https://doi.org/10.1002/wfs2.1395>. To save time and to avoid the loss of DNA in the extraction, isolation, and quantitation process, direct PCR amplification techniques have been developed. *See, e.g.,* Sarah E. Cavanaugh & Abigail S. Bathrick, *Direct PCR Amplification of Forensic Touch and Other Challenging DNA Samples: A Review*, 32 Forensic Sci. Int’l: Genetics 40 (2018), <https://doi.org/10.1016/j.fsigen.2017.10.005>; Belinda Martin et al., *Comparison of Six Commercially Available STR Kits for Their Application to Touch DNA Using Direct PCR*, 4 Forensic Sci. Int’l: Reports 100243 (2021), <https://doi.org/10.1016/j.fsir.2021.100243>.

108. *But see* SWGDAM, Guidelines for STR Enhanced Detection Methods 2 (Oct. 6, 2014), <https://perma.cc/NYF7-R9HH> (“current . . . kits . . . entail 26 to 32 cycles of PCR amplification with 0.5 to 2 ng of DNA template”).

109. On the cellular material deposited by contact with hands, *see* Julia Burrill et al., *Corneocyte Lysis and Fragmented DNA Considerations for the Cellular Component of Forensic Touch DNA*, 51 Forensic Sci. Int’l: Genetics 102428 (2021), <https://doi.org/10.1016/j.fsigen.2020.102428>.

110. STR-CE testing of LT-DNA samples with “extra” PCR cycles is unusual in the United States. The laboratory of the Office of the Chief Medical Examiner for New York City employed a 31-cycle protocol instead of the then-established 28 cycles that generally withstood objections at the trial court level. In 2020, however, the New York Court of Appeals held in *People v. Williams*, 147 N.E.3d 1131 (N.Y. 2020), that, in light of defendant’s showing that there was a controversy

molecules of DNA commonplace in casework analysis.¹¹¹ In addition, laboratory studies have shown that PCR-related methods of whole-genome amplification can make copies of a single DNA molecule that then may yield usable STR profiles with STR-CE.¹¹²

But when PCR is used on such small samples of DNA, chance effects might result in one allele being amplified much more than another. Alleles can drop out, small peaks from unusual alleles at other loci can appear, stutter peaks tend to be larger in relation to their parent peak, and bits of extraneous DNA can contribute to the profile. Laboratories normally conduct experiments with samples at successively lower concentrations not only to determine an “analytical threshold” at which a true peak reliably can be distinguished from low-level “noise” but also to establish a higher “stochastic threshold” at which the sporadic effects become apparent.¹¹³ Laboratory protocols for acquiring further data and interpreting electropherograms from samples that exhibit stochastic effects vary and are discussed in connection with the interpretation of mixed DNA samples and the coming of age of probabilistic genotyping software (see section titled “Mixtures” below).

Although there are tests to estimate the quantity of DNA in a sample, whether a particular sample contains enough human DNA to allow PCR-based typing cannot always be predicted accurately, and it is not scientifically necessary to apply an arbitrary threshold as a prerequisite for testing.¹¹⁴ If a result is

among scientists on the use of the additional cycles, it was an abuse of discretion to deny his motion for a pretrial hearing on the general acceptance of the modified PCR procedure.

In response to criticism of the LT-DNA evidence in the unsuccessful prosecution of Sean Hoey for a terrorist bombing in Omagh, Northern Ireland, England briefly suspended its use of the Forensic Science Service’s 34-cycle procedure. Duncan Campbell & Vikram Dodd, *Police Suspend Use of Discredited DNA Test After Omagh Acquittal*, Guardian, Dec. 22, 2007. After it reinstated the method (with an added quantification step), the English Court of Appeal opined that “Low Template DNA can be used to obtain profiles capable of reliable interpretation if the quantity of DNA that can be analysed is above the stochastic threshold [of] between 100 and 200 picograms.” *R. v. Reed*, [2009] (CA Crim. Div.) EWCA Crim. 2698, ¶ 74; Denise Syndercombe Court, *Low Copy Number DNA: Where Next?*, 50 Med., Sci. & Law 55 (2010), <https://doi.org/10.1258/msl.2010.010015>.

111. David Moore et al., *A Comprehensive Study of Allele Drop-In over an Extended Period of Time*, 48 Forensic Sci. Int’l: Genetics 102332 (2020), <https://doi.org/10.1016/j.fsigen.2020.102332>.

112. Man Chen et al., *Comparison of CE- and MPS-based Analyses of Forensic Markers in a Single Cell After Whole Genome Amplification*, 45 Forensic Sci. Int’l: Genetics 102211 (2020) <https://doi.org/10.1016/j.fsigen.2019.102211>; Richard Jäger, *New Perspectives for Whole Genome Amplification in Forensic STR Analysis*, 23 Int’l J. Molecular Sci. 7090 (2022), <https://doi.org/10.3390/ijms23137090>; Xu Qiannan et al., *Evaluating the Effects of Whole Genome Amplification Strategies for Amplifying Trace DNA Using Capillary Electrophoresis and Massive Parallel Sequencing*, 56 Forensic Sci. Int’l: Genetics 102599 (2022) <https://doi.org/10.1016/j.fsigen.2021.102599>.

113. See John M. Butler, *Advanced Topics in Forensic DNA Typing: Interpretation* 40–44 & 93–95 (2015).

114. DNA concentration levels such as 100 or 200 picograms have been suggested to differentiate a conventional DNA profile from a low-level target profile. See *supra* note 110. However, STR-CE technology has improved since these numbers were proposed (see Davis R.L. Watkins

obtained, and if the controls (samples of known DNA and blank samples) have behaved properly, then the sample had enough DNA.¹¹⁵ This is so whether or not the label low-copy number (LCN) or low template (LT) applies.¹¹⁶

Was the Sample of Sufficient Quality?

The primary determinant of DNA quality for forensic analysis is the extent to which the long DNA molecules are intact. Within the cell nucleus, each molecule of DNA extends for millions of base pairs. Outside the cell, DNA spontaneously degrades into smaller fragments at a rate that depends on temperature, exposure to oxygen, and, most importantly, the presence of water.¹¹⁷ In dry biological samples,

et al., *Revisiting Single Cell Analysis in Forensic Science*, 11 Sci. Reps. 7054 (2021), <https://doi.org/10.1038/s41598-021-86271-6>, & authorities cited), and “[t]hresholds are often difficult to apply in a meaningful way, for example, in a sample that comprises DNA from two or more individuals, the total quantifiable does not reflect the individual contributions.” Forensic Science Regulator (U.K.), Guidance: The Interpretation of DNA Evidence (Including Low-Template DNA), FSR-G-202 § 5.2.4 (2020), <https://perma.cc/TX2Y-5SXL>.

115. Different views have been expressed on the desirability of performing replicate tests (splitting the sample into several PCR tubes) and on how to interpret the resulting profiles. The “consensus method” discards all possible alleles other than those that are detected in two or more of the resulting electropherograms. Simon Cowen et al., *An Investigation of the Robustness of the Consensus Method of Interpreting Low-Template DNA Profiles*, 5 Forensic Sci. Int’l: Genetics 400 (2011), <https://doi.org/10.1016/j.fsigen.2010.08.010>. As statistical methods for modeling the sources of variability have advanced, replicate testing has fallen out of favor. See, e.g., David J. Balding, *Evaluation of Mixed-source, Low-template DNA Profiles in Forensic Science*, 110 Proc. Nat’l Acad. Sci. 12241 (2013), <https://doi.org/10.1073/pnas.1219739110>; Forensic Science Regulator, *supra* note 114, § 11.1.4. Replicate profiling of small samples continues to be discussed. E.g., Forensic Science Regulator, *supra*, § 12; Simone Gittelsohn et al., *Low-template DNA: A Single DNA Analysis or Two Replicates?*, 264 Forensic Sci. Int’l 139 (2016), <https://doi.org/10.1016/j.forsciint.2016.04.012>; SWGDAM, *supra* note 108, § 4.1.

116. The phrase “low copy number” originally was coined for the Forensic Science Service’s method that used 34 PCR cycles. It has caused confusion. E.g., *United States v. McCluskey*, 954 F. Supp. 2d 1224, 1278 (D.N.M. 2013) (skirting the debate between the parties on whether LCN refers to modifications in the testing or to the quantity of DNA by focusing on “whether the Government’s DNA results in this case are reliable and admissible” rather than “establish[ing] a definition of ‘LCN testing’”); *United States v. Davis*, 602 F. Supp. 2d 658 (D. Md. 2009) (avoiding “making a finding with regard to the dueling definitions of LCN testing advocated by the parties”); *State v. Bigger*, 254 P.3d 1142, 1151–52 (Ariz. Ct. App. 2011) (“We need not determine the proper definition for low copy number, or whether the DNA testing used in this case is labeled properly as LCN.”). The more generic term “low template” (LT) DNA analysis includes the use of longer injection times for CE and sample concentration methods. Forensic Science Regulator, *supra* note 114, § 5.2.1. SWGDAM, *supra* note 108, introduced a more complicated definition.

117. Other forms of chemical alteration to DNA are well studied, both for their intrinsic interest and because chemical changes in DNA are a contributing factor in the development of cancers in living cells. Some forms of DNA modification, such as those produced by exposure to ultraviolet radiation, inhibit the amplification step in PCR-based tests, whereas other chemical

protected from air and not exposed to temperature extremes, DNA degrades very slowly. STR testing has proved effective with old and badly degraded material such as the remains of the Tsar Nicholas family (buried in 1918 and recovered in 1991).¹¹⁸

The extent to which degradation affects a PCR-based test depends on the size of the DNA segment to be amplified. For example, in a sample in which the bulk of the DNA has been degraded to fragments well under 1,000 base pairs in length, it may be possible to amplify a 100–base-pair sequence, but not a 1,000–base-pair target. Consequently, the shorter alleles may be detected in a highly degraded sample, but the larger ones may be missed. Inasmuch as the size differences among STR alleles at a locus are quite small (typically no more than 50 base pairs), “locus dropout” is more common than the dropout of only one allele at a heterozygous locus.¹¹⁹

DNA can be exposed to a great variety of environmental insults without any effect on its capacity to be typed correctly. Exposure studies have shown that contact with a variety of surfaces, both clean and dirty, and with gasoline, motor oil, acids, and alkalis either have no effect on DNA typing or, at worst, render the DNA untypable.¹²⁰

Although contamination with microbes generally does little more than degrade the human DNA, other problems sometimes can occur. Therefore, the validation of DNA typing systems should include tests for interference with a variety of microbes to see if artifacts occur. If artifacts are observed, then control tests should be applied to distinguish between the artifactual and the true results.

modifications appear to have no effect. Cydne L. Holt et al., *TWGDAM Validation of AmpFISTR PCR Amplification Kits for Forensic DNA Casework*, 47 J. Forensic Sci. 66 (2002), <https://doi.org/10.1520/jfs15206j>; George F. Sensabaugh & Cecilia von Beroldingen, *The Polymerase Chain Reaction: Application to the Analysis of Biological Evidence*, in *Forensic DNA Technology* 63 (Mark A. Farley & James J. Harrington eds., 1991).

118. Peter Gill et al., *Identification of the Remains of the Romanov Family by DNA Analysis*, 6 *Nature Genetics* 130 (1994), <https://doi.org/10.1038/ng0294-130>.

119. In particular, in cases involving severe degradation, loci yielding products greater than 200 base pairs may not be detected. Holt et al., *supra* note 117. Special primers and very short STRs give better results with extremely degraded samples. See Michael D. Coble & John M. Butler, *Characterization of New MiniSTR Loci to Aid Analysis of Degraded DNA*, 50 J. Forensic Sci. 43 (2005), <https://doi.org/10.1520/jfs2004216>. Sequencing methods can be even more effective for such samples. See *supra* section titled “SNPs as identification loci.”

120. Holt et al., *supra* note 117. Most of the effects of environmental insult readily can be accounted for in terms of basic DNA chemistry. For example, some agents produce degradation or damaging chemical modifications. Other environmental contaminants inhibit restriction enzymes or PCR. (This effect sometimes can be reversed by cleaning the DNA extract to remove the inhibitor.) But environmental insult does not result in the selective loss of an allele at a locus or in the creation of a new allele at that locus.

Laboratory Performance

What Quality Control and Assurance Measures Are in Place?

DNA profiling of suitable samples is valid and reliable when properly conducted, but confidence in a particular result also depends on the quality control and quality assurance procedures in the laboratory. Quality control refers to measures to help ensure that DNA-typing results and their interpretation meet a specified standard of quality. Quality assurance refers to monitoring, verifying, and documenting laboratory performance. A quality assurance program helps demonstrate that a laboratory is meeting its quality control objectives.¹²¹

Professional bodies within forensic science have described general procedures for quality assurance. Guidelines for DNA analysis have been prepared by FBI-appointed groups (the current incarnation is known as the Scientific Working Group on DNA Analysis Methods [SWGDM]),¹²² the federally supported Organization of Scientific Area Committees for Forensic Science (OSAC),¹²³ and private Standards Developing Organizations (SDOs).¹²⁴

121. For a review of the history of quality assurance in forensic DNA testing, see Joseph L. Peterson et al., *The Feasibility of External Blind DNA Proficiency Testing. I. Background and Findings*, 48 J. Forensic Sci. 21, 22 (2003), <https://doi.org/10.4324/9780203380826>.

122. The FBI established the Technical Working Group on DNA Analysis Methods (TWGDAM) in 1988 to develop standards. The DNA Identification Act of 1994, 34 U.S.C. §§ 12591(a)–(c), 14131(a) & (c), created a DNA Advisory Board (DAB) to assist in promulgating quality assurance standards, but the legislation allowed the DAB to expire after five years (unless extended by the director of the FBI). 34 U.S.C. § 12591(b)(4). TWGDAM functioned under the DAB, 34 U.S.C. § 12591(a)(4), and was renamed the Scientific Working Group on DNA Analysis Methods (SWGDM) in 1999. When the FBI allowed DAB to expire, SWGDAM replaced it. See Norah Rudin & Keith Inman, *An Introduction to Forensic DNA Analysis* 180 (2d ed. 2002); Paul C. Giannelli, *Regulating Crime Laboratories: The Impact of DNA Evidence*, 15 J.L. & Pol’y 59, 82–83 (2007). SWGDAM characterizes the FBI quality-assurance standards as establishing minimum requirements that it supplements with more detailed, noncompulsory guidelines. SWGDAM, Frequently Asked Questions, <https://perma.cc/Z9AR-FG6A>, last visited Sept. 9, 2025.

123. The Department of Commerce’s National Institute of Standards and Technology (NIST) supports OSAC’s standards-writing efforts and maintains “a repository of selected published and proposed standards for forensic science . . . to promote valid, reliable and reproducible forensic results.” NIST, OSAC Registry (updated Dec. 22, 2023), <https://perma.cc/M9LS-96UN>. The OSAC Registry maintains that “the standards on this Registry have undergone a technical and quality review process that actively encourages feedback from forensic science practitioners, research scientists, human factors experts, statisticians, legal experts, and the public.” *Id.* The standards are voluntary, and some have been criticized as vague or trivial. Geoffrey Stewart Morrison et al., *Vacuous Standards—Subversion of the OSAC Standards-development Process*, 2 Forensic Sci. Int’l: Synergy 206 (2020), <https://doi.org/10.1016/j.fs SYN.2020.06.005>. The rigor of the actual review process, particularly as it involves independent scientific and statistical review, also has been questioned. See Kaye et al., *supra* note 5, § 12.5.1, at 680–82; Maneka Sinha, *Radically Reimagining Forensic Evidence*, 73 Ala. L. Rev. 879 (2022).

124. The major SDO that publishes voluntary standards for forensic DNA analysis is the Academy Standards Board (ASB) of the American Academy of Forensic Sciences. AAFS, Academy

A small number of states require forensic DNA laboratories to be accredited,¹²⁵ and federal law requires accreditation or other safeguards for laboratories that receive certain federal funds¹²⁶ or participate in the national DNA database system.¹²⁷

Documentation

Quality assurance guidelines normally call for laboratories to document laboratory organization and management, personnel qualifications and training, facilities, evidence control procedures, validation of methods and procedures, analytical procedures, equipment calibration and maintenance, standards for case documentation and report writing, procedures for reviewing case files and testimony, proficiency testing, corrective actions, audits, safety programs, and review of subcontractors.

Of course, maintaining documentation and records alone does not guarantee the correctness of results obtained in any particular case. Measurement error can be estimated but not eliminated, and process errors in analysis or

Standards Board (2022), <https://perma.cc/Y8N4-A4FU>. Inasmuch as the dominant voice in both OSAC and the SDOs that promulgate forensic-science standards is the practicing forensic-science community, a consensus standard does not necessarily reflect a consensus of experts in the broader scientific community. Within OSAC and the ASB, agreement of two-thirds or more of the eligible voters constitutes a consensus. The votes are not published.

125. Joseph Peterson & Matthew Hickman, *Forensic Science Practice in the United States*, in *The Global Practice of Forensic Science* 301, 331 (Douglas H. Ubelaker ed., 2015). New York was the first state to impose this requirement. N.Y. Exec. Law § 995-b (McKinney 2006) (requiring accreditation by the state Forensic Science Commission). Only one state, Texas, requires its forensic analysts to be licensed. Texas Forensic Sci. Comm'n, Forensic Analyst Licensing Program, <https://perma.cc/6UDT-NU3W>, last visited Sept. 9, 2025.

126. The Justice for All Act, enacted in 2004, required DNA labs to be accredited within two years “by a nonprofit professional association of persons actively involved in forensic science that is nationally recognized within the forensic science community” and to “undergo external audits, not less than once every 2 years, that demonstrate compliance with standards established by the Director of the Federal Bureau of Investigation.” 34 U.S.C. § 12592(b)(2)(A). The 2004 Act also requires applicants for federal funds for forensic laboratories to certify that the laboratories use “generally accepted laboratory practices and procedures, established by accrediting organizations or appropriate certifying bodies,” 34 U.S.C. § 10562(2), and that “a government entity exists and an appropriate process is in place to conduct independent external investigations into allegations of serious negligence or misconduct substantially affecting the integrity of the forensic results committed by employees or contractors of any forensic laboratory system, medical examiner’s office, coroner’s office, law enforcement storage facility, or medical facility in the State that will receive a portion of the grant amount.” *Id.* § 10562(4).

127. See 34 U.S.C. § 12592(b)(2) (requiring that records in the database come from laboratories that “have been accredited by a nonprofit professional association . . . and . . . undergo external audits, not less than once every 2 years [and] that demonstrate compliance with standards established by the Director of the Federal Bureau of Investigation”).

interpretation can occur as a result of a deviation from an established procedure, an analyst misjudgment, or an accident. Although administrative and technical review procedures within a laboratory should be designed to detect errors before a report is issued, it is always possible that some incorrect result will slip through.¹²⁸ Accordingly, determination that a laboratory maintains a strong quality-assurance program, either on paper or in practice, does not guarantee accuracy in all cases.¹²⁹

Validation

The validation of procedures is central to quality assurance. Developmental validation is undertaken to determine the applicability of a new test or equipment to crime-scene samples; it defines conditions that give accurate results and identifies the limitations of the procedure.¹³⁰ For example, a new genetic locus being considered for use in forensic analysis will be tested in both fresh samples and in samples typical of those found at crime scenes. The validation would include samples originating from different tissues, samples containing degraded DNA, samples contaminated with microbes, samples containing DNA mixtures, and so on. A valid typing method usually will report a genotype that is in a test sample and usually will not report a genotype that is not in the test sample.¹³¹ Developmental validation of a new set of loci also includes the generation of population databases and the testing of alleles for statistical independence. Developmental validation normally results in publication in the scientific literature and reflects scientific norms for theoretical and

128. See U.S. Dep't of Just., Office of the Inspector Gen., *The FBI DNA Laboratory: A Review of Protocol and Practice Vulnerability* (2004), available at <https://oig.justice.gov/reports/fbi-dna-laboratory-review-protocol-and-practice-vulnerabilities>.

129. The District of Columbia's independent, state-of-the-art Department of Forensic Sciences twice lost its accreditation—and each of its first two directors—following complaints from prosecutors, first about DNA mixture analyses and later about firearms-toolmark examinations. See Keith L. Alexander & Julie Zauzmer, *Director of D.C.'s Embattled DNA Lab Resigns After Suspension of Testing*, Wash. Post, May 1, 2015, at B1, <https://perma.cc/Z4FM-EWWB>; Editorial, *The D.C. Crime Lab Is in Trouble—Again*, Wash. Post, June 4, 2021, at A20.

130. See *supra* section titled “What Establishes That a Genetic System Is Valid for Identification?”

131. No test is *always* correct. For a set of test samples, all of which contain a particular genotype, a test might be positive for that genotype in, say, 96% of them. This proportion is the observed “sensitivity” in the test set. For a set of test samples, none of which contain the genotype, the test might be negative for that genotype in, say, 99% of them. This proportion is the observed “specificity” for that genotype in the test set. A perfect test would have 100% sensitivity and 100% specificity in all properly conducted experiments. For further discussion of the validity of binary (present-or-absent) testing, see Kaye & Stern, *supra* note 12.

empirical support of claims for methods,¹³² but the empirical aspects for validating a new procedure can be accomplished in multiple laboratories well ahead of publication.¹³³

So-called internal validation, on the other hand, involves the capacity of a specific laboratory to analyze the new loci or use the new methodology or equipment.¹³⁴ The laboratory should verify that it can reliably perform an established procedure that already has undergone developmental validation.¹³⁵ In particular, before adopting a new procedure, the laboratory should verify that its analysts can obtain correct results with samples that are representative of casework.

Proficiency testing

Proficiency testing in forensic genetic testing is designed to ascertain whether an analyst can correctly determine genetic types in a sample whose origin is unknown to the analyst but is known to a tester. Proficiency is demonstrated by making correct determinations in repeated trials. The laboratory also can be tested to verify that it correctly computes random-match probabilities or likelihood ratios.

An internal proficiency trial is devised and conducted within a laboratory.¹³⁶ One person in the laboratory prepares the sample and then administers the test to another person in the laboratory. In an external trial, the test sample originates

132. As such, validation is not easily reduced to standards defining the research that needs to be undertaken, and validation standards often seem open textured. *See supra* note 103.

133. SWGDAM “encourage[s]” but does not require publication of developmental validity studies in scientific venues. SWGDAM, *supra* note 103, § 2.2.1.2. Whether the validity required for admission of scientific evidence can be established in the absence of a robust set of scientific publications is a legal rather than a strictly scientific question. However, it has been argued that scientific norms demand multiple studies with confirmation coming from separate research groups rather than only from the developer of a method. President’s Council of Advisors on Sci. & Tech., *supra* note 105, at 80.

134. This line between developmental and internal validation is not always clear. Successful internal validation efforts add to the body of knowledge demonstrating that an analytical method performs as it should and thus enhances or extends what might be seen as a weak but encouraging previous validation. Furthermore, “internal validation” sometimes designates finding information needed to fine-tune a method. That function is better referred to as calibration.

135. Validation builds on the accumulated body of knowledge and experience. Thus, some aspects of validation testing need be repeated only to the extent required to verify that previously established principles apply.

136. Some forensic-science organizations do not use the term “proficiency testing” for tests of examiner performance developed or conducted within a single laboratory. In their view, proficiency testing requires interlaboratory comparisons. OSAC, OSAC Preferred Terms 2 (Aug. 2022), <https://perma.cc/KJC4-6HWE>.

from outside the laboratory—from another laboratory, a commercial vendor, or a regulatory agency. In a declared (or open) proficiency trial, the analyst knows the sample is a proficiency sample. The DNA Identification Act of 1994 required proficiency testing for analysts in the FBI as well as those in laboratories participating in the national database or receiving federal funding,¹³⁷ but these matters are now left to the discretion of the director of the FBI.¹³⁸ The standards of accrediting bodies typically call for periodic open, external proficiency testing.¹³⁹

In a blind or, more properly, “full blind” trial, the sample is submitted so that the analyst does not recognize it as a proficiency sample. A full-blind trial provides a better indication of proficiency because it ensures that the analyst will not give the trial sample any special attention, and it tests more steps in the laboratory’s processing of samples. However, full-blind proficiency trials entail considerably more organizational effort and expense than open proficiency trials. Obviously, the “evidence” samples prepared for the trial have to be sufficiently realistic that the laboratory does not suspect the legitimacy of the submission. A police agency and prosecutor’s office have to submit the “evidence” and respond to laboratory inquiries with information about the “case.” Finally, the genetic profile from a proficiency test must not be entered into regional and national databases. Consequently, although some forensic DNA laboratories participate in full-blind testing, they are not required to do so.¹⁴⁰

137. Pub. L. 103–22 § 210304(b)(2) (codified as amended 34 U.S.C. § 12592(b)(2)(A)(ii)) (requiring external proficiency testing of laboratories for participation in the national database); *id.* § 210305(a)(1)(A) (same for FBI examiners).

138. 34 U.S.C. § 12592(b)(2)(A)(ii) (requiring “external audits, not less than once every 2 years, that demonstrate compliance with standards established by the Director of the Federal Bureau of Investigation”). The standards require periodic external proficiency testing. FBI, Quality Assurance Standards for Forensic DNA Testing Laboratories § 13.1 (2020), <https://le.fbi.gov/file-repository/forensic-qas-070120.pdf>.

139. See Peterson et al., *supra* note 121, at 24 (describing the ASCL-LAB standards). Certification by the American Board of Criminalistics as a specialist in forensic-biology DNA analysis requires one proficiency trial per year. Accredited laboratories must maintain records documenting compliance with required proficiency-test standards.

140. However, laboratory management can blind a particular analyst within the laboratory by injecting disguised “cases” into the usual flow of cases. Section 210303(c) of the DNA Identification Act of 1994, Pub. L. 103–22, 108 Stat. 2069, required the director of the National Institute of Justice to report to Congress on the feasibility of establishing an external blind-proficiency-testing program for DNA laboratories. A National Forensic DNA Review Panel advised the director that “blind proficiency testing is possible, but fraught with problems” of the kind listed above. Peterson et al., *supra* note 121, at 30. It “recommended that a blind proficiency testing program be deferred for now until it is more clear how well implementation of the first two recommendations [the promulgation of guidelines for accreditation, quality assurance, and external audits of casework] are serving the same purposes as blind proficiency testing.” *Id.*

How Are Samples Created and Handled?

Sample mishandling, mislabeling, or contamination, whether in the field or in the laboratory, is more likely to compromise a DNA analysis than is an error in genetic typing. For example, a sample mix-up due to mislabeling reference blood samples taken at the hospital could lead to an incorrect association of crime-scene samples to a reference individual or to incorrect exclusions. Similarly, packaging two items with wet blood stains into the same bag could result in a transfer of stains between the items, rendering it difficult or impossible to determine whose blood was originally on each item. Contamination in the laboratory may result in artifactual typing results or in the incorrect attribution of a DNA profile to an individual or to an item of evidence. Procedures should be prescribed and implemented to guard against such error.

Mislabeling or mishandling can occur when biological material is collected in the field, when it is transferred to the laboratory, when it is in the analysis stream in the laboratory, when the analytical results are recorded, or when the recorded results are transcribed into a report. Mislabeling and mishandling can happen with any kind of physical evidence and are of great concern in all fields of forensic science. Checkpoints should be established to detect mislabeling and mishandling along the line of evidence flow. Investigative agencies should have guidelines for evidence collection and labeling so that a chain of custody is maintained. Similarly, there should be guidelines, produced with input from the laboratory, for handling biological evidence in the field.

Professional guidelines and recommendations require documented procedures to ensure sample integrity and to avoid sample mix-ups, labeling errors, recording errors, and the like. They also mandate case review to identify inadvertent errors before a final report is released. Finally, laboratories must retain, when feasible, portions of the crime-scene samples and extracts to allow reanalysis.¹⁴¹ However, retention is not always possible. For example, retention of original items is not to be expected when the items are large or immobile (for example, a wall or sidewalk). In such situations, a swabbing, scraping, or cutting of the stain from the item would typically be collected and retained. There also are situations where the sample is so small that it will be consumed in the analysis.

141. See FBI, *supra* note 138, § 7.4.1. Furthermore, failure to preserve potentially exculpatory evidence has been treated as a denial of due process and grounds for suppression. *People v. Nation*, 604 P.2d 1051, 1054–55 (Cal. 1980). In *Arizona v. Youngblood*, 488 U.S. 51 (1988), however, the Supreme Court held that a police agency's failure to preserve evidence not known to be exculpatory does not constitute a denial of due process unless "bad faith" can be shown. Ironically, DNA testing that was not available at Youngblood's trial established that he had been falsely convicted. Maurice Possley, *DNA Exonerates Inmate Who Lost Key Test Case: Prosecutors Ruined Evidence in Original Trial*, Chi. Trib., Aug. 10, 2000, at 6, <https://perma.cc/F78E-E8K7>.

Assuming that appropriate chain-of-custody and evidence-handling protocols are in place, the critical question is whether there are deviations in a particular case. Answering this question may require a review of the total case documentation as well as the laboratory findings. In addition, when retesting original evidence items or the material extracted from them is possible, it can be used to guard against error from mislabeling and mishandling. Should mislabeling or mishandling have occurred, reanalysis of the original sample and the intermediate extracts may detect not only the fact of the error but also the point at which it occurred.¹⁴²

Contamination describes any situation in which foreign material is mixed with a sample of DNA.¹⁴³ As noted in the section titled “Was the Sample of Sufficient Quality?” above, contamination by nonbiological materials, such as gasoline or grit, can cause test failures, but it is not a source of genetic typing errors. Similarly, contamination with nonhuman biological materials, such as bacteria, fungi, or plant materials, is generally not a problem. These contaminants may accelerate DNA degradation, but they do not generate spurious human genetic types.

The contamination of greatest concern results from the addition of human DNA. This sort of contamination can occur in three ways. First, the crime-scene samples, by their nature, may contain a mixture of fluids or tissues from different individuals. Examples include vaginal swabs collected as sexual-assault evidence and bloodstain evidence from scenes where several individuals shed blood. The analysis of such mixtures is the subject of the “Mixtures” section below.

Second, the crime-scene samples may be inadvertently contaminated in the course of sample handling in the field or in the laboratory. “Elimination databases” of DNA profiles from laboratory personnel as well as police officers and emergency responders who are present at crime scenes can be used to see

142. Of course, retesting cannot correct all errors that result from mishandling of samples, but it is possible in some cases to detect mislabeling at the point of sample collection if the genetic-typing results on a particular sample are inconsistent with an otherwise consistent reconstruction of events. For example, a mislabeling of husband and wife samples in a paternity case might result in an apparent maternal exclusion, a very unlikely event. The possibility of mislabeling could be confirmed by testing the samples for gender and ultimately verified by taking new samples from each party under better-controlled conditions.

143. *Cf.* Forensic Science Regulator (U.K.), Forensic Science Regulator Guidance: Contamination Controls—Scene of Crime § 1.1.1, at 4 (2023) (FSR-GUI-0016), <https://perma.cc/Q4D2-MGJU> (“For the purposes of this guidance, contamination is defined as ‘the undesirable introduction of DNA, or biological material containing DNA, to an item/exhibit recovered from an incident scene which is to be recovered/analysed and before a controlled forensic process is started’.” This is distinct from the adventitious transfer of biological material to an exhibit that can also occur, usually prior to the exhibit or sample being recovered and before investigative agencies have intervened, this is often referred to as ‘background DNA.’”); Forensic Science Regulator (U.K.), Forensic Science Regulator Guidance: DNA Contamination Controls—Laboratory § 1.1.1, at 5 (2023), FSR-GUI-0018, <https://perma.cc/GU96-M769> (same).

if DNA from these individuals might be in the recovered samples.¹⁴⁴ Inadvertent contamination of crime-scene DNA with DNA from a reference sample could lead to a false inclusion. Likewise, with low-template DNA samples, secondary transfer—in which some DNA molecules from one person are present for a time on another person and then are deposited in the crime-scene sample—could be falsely incriminating.¹⁴⁵ Research into mechanisms of DNA transfer and DNA- and RNA-based testing that might help clarify the origin of the trace DNA is discussed in the section titled “DNA and RNA Typing of Bodily Fluids or Tissues” below.

Third, carryover contamination in PCR-based typing can occur if the amplification products of one typing reaction are carried over into the reaction mix for a subsequent PCR reaction. If the carryover products are present in sufficient quantity, they could be preferentially amplified over the target DNA. To protect against carryover contamination, the primary strategy is to keep PCR products away from sample materials and test reagents by having separate work areas for pre-PCR and post-PCR sample handling, by preparing samples in controlled-airflow biological safety hoods, by using dedicated equipment (such as pipettes) for each of the various stages of sample analysis, by decontaminating work areas after use (usually by wiping down or by irradiating with ultraviolet light), and by having a one-way flow of sample from the pre-PCR

144. Martine Lapointe et al., *Leading-edge Forensic DNA Analyses and the Necessity of Including Crime Scene Investigators, Police Officers and Technicians in a DNA Elimination Database*, 19 *Forensic Sci. Int'l: Genetics* 50 (2015), <https://doi.org/10.1016/j.fsigen.2015.06.002>; OSAC, *Best Practice Recommendations for the Management and Use of Quality Assurance DNA Elimination Databases in Forensic DNA Analysis* (Apr. 6, 2021) (2020-N-0007). However, the Genetic Information and Nondiscrimination Act of 2008, 42 U.S.C. § 2000ff et seq., prohibits employers from requesting “genetic information” from employees. An exception, in § 202(b)(6) of the act, applies “where the employer conducts DNA analysis for law enforcement purposes as a forensic laboratory or for purposes of human remains identification, and requests or requires genetic information of such employer’s employees . . . for quality control to detect sample contamination.” 42 U.S.C. § 2000ff-1(b)(6). Non-laboratory police employees have claimed that because they fall outside this exception, the statute protects them from being required to provide elimination samples. In response, it can be argued that the statute’s expansive definition of “genetic test” should be read in light of its main purpose as an anti-discrimination law concerned with health-related genetic tests. David H. Kaye, *GINA’s Genotypes*, 108 *Mich. L. Rev. First Impressions* 51 (2010). This intent- or purpose-based approach to the statute was rejected in *Lowe v. Atlas Logistics Group Retail Services (Atlanta)*, 102 F. Supp. 3d 1360 (N.D. Ga. 2015) (construing GINA on the basis of its plain text as prohibiting an employer’s request for a DNA sample for STR profiling to find “a mystery employee [who was] habitually defecating in one of its warehouses”); see also David H. Kaye, *EEOC Stays Mum on GINA*, *Forensic Sci., Stat. & L.* (Jan. 22, 2013), <https://perma.cc/QS9Z-XAYC>.

145. See Peter Gill, *Misleading DNA Evidence: A Guide for Scientists, Judges, and Lawyers* (2014).

to post-PCR work areas.¹⁴⁶ Additional protocols are used to detect carryover contamination.¹⁴⁷

In the end, whether a laboratory has conducted proper tests depends both on the general standard of practice and on the questions posed in the particular case. There is no universal checklist, but the selection of tests and the adherence to the correct test procedures can be reviewed by experts and by reference to professional standards.

Inference, Statistics, and Population Genetics in Human Nuclear DNA Testing

What Constitutes a Match? An Exclusion?

The results of DNA testing can be presented in various ways. With discrete allele systems, such as STRs, it is natural—but not essential—to speak of “matching” and “nonmatching” profiles. If the profile obtained from the single-source biological sample taken from the crime scene or the victim (the trace-evidence sample) clearly differs from the DNA in a sample known to come from a particular individual (the reference sample), that individual is excluded as a possible source. At the other extreme, two multilocus genotypes can be clearly identical. In these cases, the DNA evidence typically is quite incriminating,¹⁴⁸ but even when two samples have the same genotype, there is a chance that the trace-evidence sample came not from the defendant, but from another individual who has the same genotype. As indicated in the section above titled “A Brief History

146. SWGDAM, Contamination Prevention and Detection Guidelines for Forensic DNA Laboratories (Jan. 12, 2017), <https://perma.cc/R2SD-BJ8A>.

147. Standard protocols include the amplification of blank control samples—those to which no DNA has been added. If carryover contaminants have found their way into the reagents or sample tubes, these will be detected as amplification products. Outbreaks of carryover contamination can also be recognized by monitoring test results. Detection of an unexpected and persistent genetic profile in different samples indicates a contamination problem. When contamination outbreaks are detected, corrective actions should be taken, and both the outbreak and the corrective action should be documented. Human Forensic Biology Subcomm., Org. of Sci. Area Committees for Forensic Sci., Standard for Interpreting, Comparing and Reporting DNA Test Results Associated with Failed Controls and Contamination Events (OSAC 2020-S-0004, Ver. 2.0, May 2021).

148. Whether being the source of the forensic sample is incriminating and whether someone else being the source is exculpatory depends on the circumstances. For example, a suspect who might have committed the offense without leaving the trace-evidence sample still could be guilty. In a rape case with several rapists, a semen stain could fail to incriminate one assailant because insufficient semen from that individual is present in the sample.

of DNA Evidence,” this complication has produced extensive arguments over the statistical procedures for assessing this possibility.

Some cases lie between the poles of a clear inclusion or exclusion. Since the earliest days of RFLP-VNTR testing, concerns have been expressed about subjective aspects of procedures that leave room for “observer effects”¹⁴⁹ in interpreting data.¹⁵⁰ When the trace-evidence sample is small and extremely degraded, STR profiling can be afflicted with allelic drop-in, drop-out, and other complications.¹⁵¹ Analysts operating within the framework of reporting that alleles are categorically present or absent then must make judgments as to whether true peaks are missing and whether spurious peaks are present.¹⁵² Experts then might disagree about whether a suspect is included or excluded—or whether any conclusion can be drawn.¹⁵³

What Hypotheses Can Be Formulated About the Source?

In DNA identification cases, the laboratory analyst reports that the sample of DNA from the crime scene and a sample from the defendant have the same genotype. There could be an innocent explanation for the defendant’s DNA being at the crime scene,¹⁵⁴ but the matching DNA might not even be the defendant’s. One

149. See generally D. Michael Risinger et al., *The Daubert/Kumho Implications of Observer Effects in Forensic Science: Hidden Problems of Expectation and Suggestion*, 90 Calif. L. Rev. 1 (2002).

150. E.g., William C. Thompson & Simon Ford, *The Meaning of a Match: Sources of Ambiguity in the Interpretation of DNA Prints*, in *Forensic DNA Technology* (Mark J. Farley & James J. Harrington eds., 1990).

151. See *supra* section titled “Sample Collection and Contamination.”

152. Commentators have proposed that the analyst determine the profile of a trace-evidence sample before knowing the profile of any suspects. E.g., Itiel E. Dror & Jeff Kukucka, *Linear Sequential Unmasking-Expanded (LSU-E): A General Approach for Improving Decision Making As Well As Minimizing Noise and Bias*, 3 *Forensic Sci. Int’l: Synergy* 100161 (2021), <https://doi.org/10.1016/j.fsisy.2021.100161>; Dan E. Krane et al., *Sequential Unmasking: A Means of Minimizing Observer Effects in Forensic DNA Interpretation*, 53 *J. Forensic Sci.* 1006 (2008), <https://doi.org/10.1111/j.1556-4029.2008.00787.x>.

153. See, e.g., *State v. Murray*, 174 P.3d 407, 417–18 (Kan. 2008) (inconclusive results were presented as consistent with the defendant’s blood).

154. For example, the two samples would match if someone framed the defendant by putting a sample of defendant’s DNA at the crime scene or in the container of DNA thought to have come from the crime scene. Or, the defendant’s DNA might have made its way to the crime-scene sample inadvertently, as is thought to have happened when a homeless man’s DNA was found on the fingernails of a multi-millionaire found dead in his home after a robbery. After a public defender established that the defendant was in a hospital when the murder occurred, police and prosecutors concluded that his DNA probably had been transferred by paramedics who treated the defendant three hours before they applied a pulse oximeter to the victim’s finger. Katie Worth, *Framed for Murder by His Own DNA*, *Frontline*, Apr. 19, 2018, <https://perma.cc/C5ST-QWNS>.

possibility is laboratory error—the genotypes are not actually the same even though the expert thinks that they are. This situation could arise from mistakes in labeling or handling samples or from cross-contamination of the samples. Another alternative is that the genotypes are truly identical but the forensic sample came from another individual. In general, the true source might be a close relative of the defendant¹⁵⁵ or an unrelated person who just happens to have the same profile as the defendant. The former hypothesis we shall refer to as kinship, and the latter as coincidence. To infer that the defendant is the source of the trace-evidence DNA, one must reject these alternative hypotheses of laboratory error, kinship, and coincidence. Table 1 summarizes the logical possibilities.¹⁵⁶

Table 1. Hypotheses that Might Explain a Match Between Defendant’s DNA and DNA at a Crime Scene

IDENTITY	
Defendant is source	Same genotype, defendant’s DNA at crime scene
NONIDENTITY	
Laboratory error	Different genotypes mistakenly found to be the same
Kinship	Same genotype, relative’s DNA at crime scene
Coincidence	Same genotype, unrelated individual’s DNA at crime scene

If laboratory error, kinship, and coincidence are rejected as implausible, then only the hypothesis of identity remains. We turn, then, to the considerations that affect the chances of a match when the defendant is not the source of the trace evidence.

Can the Match Be Attributed to Laboratory Error?

Traditional legal and scientific procedures can help to assess the possibilities of errors in handling or analyzing the samples. Scrutinizing the chain of custody, examining the laboratory’s protocol, verifying that it adhered to that protocol, and conducting confirmatory tests (including testing by the defense) can help show that the profiles really do match. Yet, “[e]rrors happen, even in the best laboratories, and even when the analyst is certain that every precaution against error was taken.”¹⁵⁷

155. A close relative, for these purposes, would be a brother, uncle, nephew, etc. For relationships more distant than second cousins, the probability of a chance match is nearly as small as for persons of the same ethnic subgroup. Bernard Devlin & Kathryn Roeder, *DNA Profiling: Statistics and Population Genetics*, in 1 *Modern Scientific Evidence: The Law and Science of Expert Testimony* § 18–3.1.3, at 724 (David L. Faigman et al. eds., 1997) (section not included in later editions).

156. Cf. Newton E. Morton, *The Forensic DNA Endgame*, 37 *Jurimetrics J.* 477, 480 tbl. 1 (1997).

157. NRC I, *supra* note 6, at 89; see also Ate Kloosterman et al., *Error Rates in Forensic DNA Analysis: Definition, Numbers, Impact and Communication*, 12 *Forensic Sci. Int’l: Genetics* 775 (2014), <https://doi.org/10.1016/j.fsigen.2014.04.014>.

Some commentary proposes using the proportion of false positives that the particular examiner or laboratory has experienced in blind proficiency tests, or the rate of false positives on proficiency tests averaged across all laboratories, to estimate the probability of a false inclusion in the case at bar.¹⁵⁸ Other commentators stress that there are too few proficiency tests to estimate averages accurately and question the application of an historical industry-wide error rate to a particular laboratory at a later time.¹⁵⁹ Proponents of using averages from proficiency tests reply that this information at least can be used to provide an upper bound on the false-positive laboratory-error probability.¹⁶⁰

Could a Close Relative Be the Source?

With enough loci to test, all individuals except some identical twins should be distinguishable. With the STR-CE technology in general use and with especially small quantities of DNA, however, this ideal is not always attainable. A thorough investigation might extend to all known relatives, but this is not always feasible, and there is always the chance that some unknown relatives are in the suspect population. Formulas are available for computing the probability that any person with a specified degree of kinship to the defendant also possesses the incriminating genotype.¹⁶¹ For example, the probability that an untested brother (or sister) would match at four loci (with alleles that each occur in 10% of the population) is about

158. E.g., Jonathan J. Koehler, *Proficiency Tests to Estimate Error Rates in the Forensic Sciences*, 12 Law, Probability & Risk 89 (2013), <https://doi.org/10.1093/lpr/mgs013> (proposing blind performance tests specifically designed to estimate error rates rather than proficiency tests for quality assurance); Jonathan J. Koehler, *Error and Exaggeration in the Presentation of DNA Evidence at Trial*, 34 *Jurimetrics J.* 21, 37–38 (1993); NRC I, *supra* note 6, at 94 (“proficiency tests provide a measure of the false-positive and false-negative rates of a laboratory”).

159. NRC II, *supra* note 7, at 85–87; Devlin & Roeder, *supra* note 155, § 18–5.3, at 744–45. Such arguments have not persuaded the proponents of estimating the probability of error from industry-wide proficiency testing. E.g., Jonathan J. Koehler, *Why DNA Likelihood Ratios Should Account for Error (Even When a National Research Council Report Says They Should Not)*, 37 *Jurimetrics J.* 425 (1997).

160. E.g., Jonathan J. Koehler, *DNA Matches and Statistics: Important Questions, Surprising Answers*, 76 *Judicature* 222, 228 (1993); Richard Lempert, *After the DNA Wars: Skirmishing with NRC II*, 37 *Jurimetrics J.* 439, 447–48, 453 (1997). For example, if there were no errors in 100 tests, a 95% confidence interval would include the possibility that the error rate could be almost as high as 3%. See NRC II, *supra* note 7, at 86 n.1. For an explanation of confidence intervals and the “zero-numerator problem,” see Kaye & Stern, *supra* note 12; see also *supra* section titled “What Are DNA Polymorphisms and How Are They Detected?”

161. John S. Buckleton et al., *Relatedness*, in *Forensic DNA Evidence Interpretation* 119 (John S. Buckleton et al. eds., 2d ed. 2016); Bruce S. Weir, *Kinship*, in *Handbook of Forensic Statistics* 265, 268–71 (David Banks et al. eds., 2021) (but also noting the uncertainties or complications in applying the formulas).

1/380;¹⁶² the probability that an aunt (or uncle) would match is about 1/100,000.¹⁶³ With more independent loci, the probabilities will plummet.

Could an Unrelated Person Be the Source?

Another rival hypothesis is coincidence: The defendant is not the source of the crime-scene DNA but happens to have the same genotype as an unrelated individual who is the true source. In principle, this hypothesis could be addressed by genotyping everyone in the suspect population. But that is rarely feasible.¹⁶⁴ Instead, an estimate of how frequently the incriminating genotype occurs in broad racial or ethnic populations is provided.¹⁶⁵ The first step is to estimate the frequencies of individual alleles from samples of various populations (often referred to as reference or population databases).¹⁶⁶ Typically, convenience samples (often from blood banks

162. For a case with conflicting calculations of the probability of an untested brother having a matching genotype, see *McDaniel v. Brown*, 558 U.S. 120 (2010) (per curiam). The correct computation is given in David H. Kaye, “False, but Highly Persuasive”: *How Wrong Were the Probability Estimates in McDaniel v. Brown?*, 108 Mich. L. Rev. First Impressions 1 (2009), https://elibrary.law.psu.edu/cgi/viewcontent.cgi?article=1059&context=fac_works; and David H. Kaye, *The Interpretation of DNA Evidence: A Case Study in Probabilities*, in Science Policy Decision-Making Educational Modules (Nat’l Academies of Sci., Engineering, and Med. Comm. on Preparing the Next Generation of Policy Makers for Science-Based Decisions ed. 2016), <https://perma.cc/D9RM-DYDK> [hereinafter Kaye, *Case Study*].

163. These figures follow from the equations in NRC II, *supra* note 7, at 113. The large discrepancy between two siblings on the one hand, and an uncle and nephew on the other, reflects the fact that the siblings have far more shared ancestry. All their genotypes are inherited through the same two parents. In contrast, a nephew and an uncle inherit from two unrelated mothers, and so will have few maternal alleles in common. As for paternal alleles, the nephew inherits not from his uncle, but from his uncle’s brother, who shares by descent only about one-half of his alleles with the uncle.

164. The suspect population normally defies enumeration, and in the typical crime where DNA evidence is found, the population of possible perpetrators is so huge that even if all of its members could be listed, they could not all be tested. As the cost of DNA profiling drops, however, it will become technically and economically feasible to have a comprehensive, population-wide DNA database that could be used to produce a list of nearly everyone whose DNA profile is consistent with the trace-evidence DNA. Whether such a system would be constitutionally and politically acceptable is another question. See David H. Kaye & Michael S. Smith, *DNA Identification Databases: Legality, Legitimacy, and the Case for Population-Wide Coverage*, 2003 Wis. L. Rev. 413. But the introduction of investigative genetic genealogy has renewed interest in the concept. J.W. Hazel et al., *Is It Time for a Universal Genetic Forensic Database?*, 362 Science 898 (2018).

165. The underlying premise is that, considering the biology of DNA markers, each possible forensic DNA genotype occurs just as often among criminal offenders as anybody else. So if we know the profile frequency in a large, relevant population group, we can use it for the unobserved set of conceivable suspects.

166. In the formative years of forensic DNA testing, defendants frequently contended that forensic databases were too small to give accurate estimates, but this argument generally proved unpersuasive. *E.g.*, *United States v. Shea*, 957 F. Supp. 331, 341–43 (D.N.H. 1997); *State v. Copeland*, 922 P.2d 1304, 1321 (Wash. 1996). To the extent that the databases are comparable to random

or paternity cases, with the “race” of the donor being self-declared or assessed by the person collecting the sample) are used to construct the databases.¹⁶⁷ Second, the estimated allele frequencies at each locus are combined into single-locus frequencies. The simplest computation posits an infinite population of individuals who choose their mates and reproduce independently of the forensic-identification alleles they possess.¹⁶⁸ Then the expected population proportion of a pair of alleles (a single-locus genotype) is essentially the product of allele frequencies.¹⁶⁹ Finally, the single-locus frequencies are multiplied together.¹⁷⁰

Because the frequencies vary across census groups (White, Black, Hispanic, Asian, and Native American), it became common to present separate estimates for these groups, at least in cases in which the race of the source of the trace evidence was unknown. For example, if a woman was abducted from a rest stop on an interstate highway and sexually assaulted, the suspect population could include people of all census categories.¹⁷¹ Random-match probabilities for all the major

samples, confidence intervals can be used to indicate the uncertainty related to sample size. Unfortunately, the meaning of a confidence interval is subtle, and the estimate commonly is misconstrued. See Kaye & Stern, *supra* note 12.

167. A few experts have testified that no meaningful conclusions can be drawn in the absence of random sampling. *E.g.*, *People v. Soto*, 88 Cal. Rptr. 2d 34 (1999); *State v. Anderson*, 881 P.2d 29, 39 (N.M. 1994). The 1996 NRC report suggests that for the purpose of estimating allele frequencies, convenience sampling should give results comparable to random sampling, and it discusses procedures for estimating the random sampling error. NRC II, *supra* note 7, at 126–27, 146–48, 186. The courts generally have rejected the argument that random samples are essential to valid or generally accepted random-match probabilities. See section titled “A Brief History of DNA Evidence” above.

168. Of course, people do not choose their mates by a lottery. “Random mating” simply indicates that the choices are uncorrelated with the specific alleles that make up the genotypes in question. A further condition of the random-mating model is that there is no systematic relationship between the particular alleles that an offspring receives and the chance that the child will transmit an allele to the next generation.

169. If the single-locus genotype is homozygous (for example, the allele from each parent is A_1), then expected proportion is just $p_1 \times p_1$, or p_1^2 , where p_1 is the proportion of all alleles in the population that are type A_1 . If the pair of alleles at a locus are distinct (call them A_1 and A_2), then the single-locus heterozygous genotype proportion is $2p_1p_2$. For example, suppose that 10% of the sperm in the gene pool of the population carry allele 1, and 50% carry allele 2. Similarly, 10% of the eggs carry A_1 , and 50% carry A_2 . (Other sperm and eggs carry other types.) With random mating, we expect $10\% \times 10\% = 1\%$ of all the fertilized eggs to be A_1A_1 , and another $50\% \times 50\% = 25\%$ to be A_2A_2 . These constitute two distinct homozygote profiles. Likewise, we expect $10\% \times 50\% = 5\%$ of the fertilized eggs to be A_1A_2 and another $50\% \times 10\% = 5\%$ to be A_2A_1 . These two configurations produce indistinguishable profiles—a peak, band, or dot for A_1 and another mark for A_2 . So the expected proportion of heterozygotes is $5\% + 5\% = 10\%$.

These proportions are known as Hardy-Weinberg proportions. Even if two populations with distinct allele frequencies are thrown together, within the limits of chance variation, random mating produces Hardy-Weinberg equilibrium in a single generation.

170. When the various single-locus genotypes are statistically independent of one another, the population is said to be in linkage equilibrium.

171. *United States v. Jakobetz*, 955 F.2d 786 (2d Cir. 1992).

groups enable the jury to see how much the estimates vary for different parts of the population even if they cannot determine which of these parts is most salient.

In cases in which the suspect population seems limited to a single major population group—for example, where the victim is sure that the assailant looked Hispanic and spoke with a Spanish accent—only the frequency estimated for that group might be presented.¹⁷² But what if the victim is mistaken or the perpetrator is of mixed ancestry? If the wrong racial or ethnic database were used to estimate a profile frequency, it would yield too small (or too large) a random-match probability. Thus, even in cases in which the evidence points to a subpopulation, there is an argument for giving the estimates for several groups.¹⁷³

A concern often raised with racial classifications involving DNA evidence is that references to race and DNA genotypes of any kind will “reify” the idea that racial categories are fundamentally biological rather than socially constructed.¹⁷⁴ There are ways to avoid the census-type labels. Allele frequency data can be collected from regions across the globe, and the spectrum of resulting profile frequency estimates could be presented in every case.¹⁷⁵ Alternatively, only the highest frequency estimate across this spectrum could be selected to supply a “conservative” estimate (that is, an overestimate). One might even create an unrealistic population by picking, for each and every allele, the frequency that is the largest in any of the groups and combining those maximum frequencies.¹⁷⁶ Ideas like these have been proposed,¹⁷⁷ and

172. Cf. *People v. Pizarro*, 12 Cal. Rptr. 2d 436, 441 (Ct. App. 1992), *after remand*, 3 Cal. Rptr. 3d 21 (Ct. App. 2003), *review denied* (Oct. 15, 2003).

173. David H. Kaye, *The Role of Race in DNA Evidence: What Experts Say, What California Courts Allow*, 37 Sw. U. L. Rev. 303 (2008).

174. E.g., Troy Duster, *Race and Reification in Science*, 307 Science 1050 (2005); section titled “Biogeographic Ancestry Testing” below.

175. The FBI’s STR allele-frequency database has samples from “African Americans, Caucasians, Southeast Hispanics, Southwestern Hispanics, Bahamians, Jamaicans, Trinidadians, Apaches, Navajos, Chamorros and Filipinos.” Tamara R. Moretti et al., *Population Data on the Expanded CODIS Core STR Loci for Eleven Populations of Significance for Forensic DNA Analyses in the United States*, 25 Forensic Sci. Int’l: Genetics 175, 175 (2016), <https://doi.org/10.1016/j.fsigen.2016.07.022>.

176. This is a rough statement of one of the controversial “ceiling” methods “strongly recommend[ed]” in NRC I, *supra* note 6, at 82–85.

177. An advisory commission appointed by the Department of Justice noted that a more refined population-genetics model (mentioned below) could be used to treat the entire United States as a single, structured population, obviating the reporting of estimates by more specific groups. Nat’l Comm’n on the Future of DNA Evidence, *The Future of DNA Evidence: Predictions of the Research and Development Working Group 5* (2000); *see also* Robert F. Oldt & Sreetharan Kanthaswamy, *Expanded CODIS STR Allele Frequencies—Evidence for the Irrelevance of Race-based DNA Databases*, 42 Legal Med. 101642 (2020), <https://doi.org/10.1016/j.jlegalmed.2019.101642> (questioning the value of racial labels when typing at 20 STR loci is successful). Another procedure that avoids racial categories is simply to give the (larger) probability of a random match to a brother or sister at the loci used in the testing. Because full siblings always are more closely related than are

some were used in cases.¹⁷⁸ However, the dominant reason given for looking to more narrowly defined populations was a perception that the model of randomly mating major racial populations was just too simple. Perhaps linguistic, religious, or other subgroups within the broader populations mate almost exclusively among themselves. If these subpopulations happen to have varying allele frequencies, then the estimates for the DNA-genotype frequencies in the broader groups could understate (or overstate) the true frequencies for the racial groups or be inapposite in a case in which the suspect population is confined to a narrow subgroup.

For a time, the likely magnitude of the effect of this “population structure” on profile-frequency estimates was hotly contested.¹⁷⁹ Today, a quantity denoted as F_{ST} or θ (theta)¹⁸⁰ often is used in formulas designed to account for population structure,¹⁸¹ however, there is a continuing discussion of what value to use for θ ,¹⁸² and some researchers have developed further refinements.¹⁸³

Frequencies, Probabilities, and Prejudice

Up to this point, we have described the estimates for genotype frequencies (often presented as random-match probabilities) that commonly are quoted in conjunction with the finding that the trace-evidence sample contains DNA of the same

randomly drawn pairs from even a narrowly defined subpopulation, this “Sib Method . . . provides a rough upper limit for the actual match probability.” Nat’l Comm’n, *supra*, at 5.

178. *E.g.*, *Brim v. State*, 695 So.2d 268 (Fla. 1997).

179. Jay D. Aronson, *Genetic Witness: Science, Law, and Controversy in the Making of DNA Profiling* (2007); Kaye, *supra* note 1, at 121–26. By the mid-1990s, the population-structure objection to admitting random-match probabilities had lost its punch. *See, e.g.*, Kaye, *supra* note 1; *supra* section titled “A Brief History of DNA Evidence.”

180. *See* David Balding & Richard A. Nichols, *DNA Profile Match Probability Calculation: How to Allow for Population Stratification, Relatedness, Database Selection and Single Bands*, 64 *Forensic Sci. Int’l* 125 (1994); Gill et al., *supra* note 42, at 29–38.

181. The recommendations of the 1996 NAS committee for computing random-match probabilities with these formulas for broad populations and particular subpopulations are summarized in the second edition of this guide. *See* Reference Manual on Scientific Evidence (2d ed. 2000).

182. *See* John Buckleton et al., *Population-specific FST Values for Forensic STR Markers: A World-wide Survey*, 23 *Forensic Sci. Int’l: Genetics* 91 (2016); Bruce S. Weir, *DNA Frequencies and Probabilities*, in *Handbook of Forensic Statistics* 251, 256–57 (David Banks et al. eds., 2021); Weir, *supra* note 161, at 266–68. The U.K. Forensic Science Regulator recommends the value of 0.03 as “sufficiently conservative that it is almost certainly favourable to defendants even allowing for alternative contributors to come from very different ethnic populations.” Forensic Science Regulator (U.K.), *Guidance: Allele Frequency Databases and Reporting Guidance for the DNA (Short Tandem Repeat) Profiling*, FSR-G-213 Issue 2, § 14.1.4 (2020); however, “in unusual cases involving small and isolated populations that may be highly differentiated from available databases,” 0.05 could be used. *Id.* § 14.1.1.

183. *See* Mark A. Jobling, *Forensic Genetics Through the Lens of Lewontin: Population Structure, Ancestry and Race*, 377 *Phil. Transactions Royal Soc’y B* 20200422 (2022) (citing three papers).

type as the defendant's. Assuming that the statistical methods meet *Daubert's* demand for scientific validity and reliability, and thus satisfy Federal Rule of Evidence 702 (or, in some states, *Frye's* requirement of general acceptance in the scientific community), a further issue can arise under Rule 403: To what extent will the presentation assist the jury in understanding the meaning of a match so that the jury can give the evidence the weight that it deserves? This question involves psychology and law, and we summarize the arguments about probative value and prejudice that have been made in litigation and in the legal and scientific literature. How the legal issue of the admissibility of any particular statistic generally should be resolved under the balancing standard of Rule 403 may turn not only on the general features of the evidence described here, but on the context and circumstances of particular cases.

Are Frequencies or Probabilities Prejudicial Because They Are So Small?

The most common form of expert testimony about matching DNA involves an explanation of how the laboratory ascertained that the defendant's DNA has the profile of the trace-evidence sample plus an estimate of the profile frequency or random-match probability.¹⁸⁴ It has been suggested, however, that jurors do not understand probabilities in general, and that infinitesimal match probabilities will so bedazzle jurors that they will not appreciate the other evidence in the case or any innocent explanations for the match.¹⁸⁵ Empirical research into this hypothesis does not clearly support the argument that jurors will overweight the probability.¹⁸⁶ The details of how the probability is presented and countered may be important, and remedies short of exclusion are available.¹⁸⁷ The practice in

184. A more flexible alternative that has widespread support among forensic DNA scientists and forensic statisticians is the likelihood ratio or Bayes' factor discussed *infra* section titled "Mixtures."

185. *Cf. Gov't of the Virgin Islands v. Byers*, 941 F. Supp. 513, 527 (D.V.I. 1996) ("Vanishingly small probabilities of a random match may tend to establish guilt in the minds of jurors and are particularly suspect.").

186. Thomas Busey et al., *Validating Strength-of-Support Conclusion Scales for Fingerprint, Footwear, and Toolmark Impressions*, 67 J. Forensic Sci. 936 (2022) ("whether jurors can understand more complex statistical terms such as likelihood ratios and random match probabilities is an empirical question with no clear answer in the literature"); see generally Kristy A. Martire, *How Well Do Lay People Comprehend Statistical Statements from Forensic Scientists*, in *Handbook of Forensic Statistics* 201 (David Banks et al. eds., 2021); David H. Kaye et al., *Statistics in the Jury Box: Do Jurors Understand Mitochondrial DNA Match Probabilities?*, 4 J. Empirical Legal Stud. 797 (2007).

187. *United States v. Graves*, 465 F. Supp. 2d 450, 458 (E.D. Pa. 2006) ("with cross-examination, proper explanations, and clarifying jury instructions, the probative value of the sneaker evidence [with RMPs of about 1/3,000] is not substantially outweighed by the danger of unfair prejudice and confusion"); *United States v. Santiago*, 156 F. Supp. 2d 145, 151 (D.P.R.

the United Kingdom is to cap reported random-match probabilities at “1 in 1 billion.”¹⁸⁸

Thus, although there once was a line of cases that excluded probability testimony in criminal matters, by the mid-1990s, no jurisdiction excluded DNA match probabilities on this basis.¹⁸⁹ The opposite argument—that relatively large random-match probabilities are prejudicial because jurors will not appreciate that they make the DNA match unimpressive—also has been advanced without much success.¹⁹⁰

Are Frequencies or Probabilities Prejudicial Because They Might Be Transposed?

A related concern is that the jury will misconstrue the random-match probability as the probability that the evidence DNA came from a random individual. The words are almost identical, but the probabilities can be different. The random-match probability is the probability that the suspect’s DNA matches the trace-evidence sample calculated on the proposition that the suspect is not the true source of that sample (and is not a close relative). The proposition is a hypothesis about who the source really is; it asserts that the true source is someone other than the suspect. But the laboratory cannot evaluate the probability of this hypothesis on the basis of

2001); NRC II, *supra* note 7, at 197 (suitable cross-examination, defense experts, and jury instructions might reduce the risk of an unwarranted sense of certainty).

188. Forensic Science Regulator, Guidance: Allele Frequency Databases and Reporting Guidance for the DNA (Short Tandem Repeat) Profiling (2014) (FSR-G-213 Issue 1, Recommendation 4, retained in Issue 2, 2020, §§ 8 & 10). In part, this limit reflects the concern that the assumptions of population genetics and statistical models do not hold exactly, reducing the probative value of much smaller numbers. It also is in the spirit of the 2023 amendment to Rule 702 intended “to emphasize that . . . [a] testifying expert’s opinion must stay within the bounds of what can be concluded by a reliable application of the expert’s basis and methodology.” Fed. R. Evid. 702 advisory committee’s note to the 2023 amendment. *But see* Tacha Hicks et al., *A Logical Framework for Forensic DNA Interpretation*, 13 *Genes* 957 (2022), <https://doi.org/10.3390/genes13060957> (Noting that when “the Forensic Science Service introduced the ten locus STR system into case-work,” the one-in-a-billion cap was used as “a temporary expedient” because it was “argued that the extent of investigations into the independence assumptions . . . was insufficient to justify the robustness of LR’s in excess of one billion”; yet “to this day, there still seems to be little in the way of large-scale between-locus dependence effects,” making the practice for the much larger “number of loci now in routine use . . . a peculiar state of affairs.”).

189. *E.g.*, *United States v. Chischilly*, 30 F.3d 1144 (9th Cir. 1994); *State v. Weeks*, 891 P.2d 477, 489 (Mont. 1995); *see also* *State v. Bauldwin*, 811 N.W.2d 267, 288 (Neb. 2012).

190. *E.g.*, *United States v. McCluskey*, 954 F. Supp. 2d 1224, 1273–75 (D.N.M. 2013); *State v. Tucker*, 920 N.W.2d 680, 688–89 (Neb. 2018). *But see* *United States v. Graves*, *supra* note 187, at 459 (“even with appropriate safeguards, the minimal probative value of the umbrella DNA evidence—in which half of the relevant population cannot be excluded as a contributor to the DNA sample—is substantially outweighed by the danger of unfair prejudice and confusion of the issues”).

the DNA evidence itself. It can only provide the probability of an event such as the suspect's DNA profile matching the crime-scene DNA profile if the true source is genetically unrelated to the suspect (or related in a specified manner). The occurrence of the match certainly is circumstantial evidence in favor of the hypothesis that the suspect is the source; however, automatically "equating source probability with random match probability" is "faulty reasoning."¹⁹¹

Sliding from the probability of the circumstantial evidence to the probability of the circumstances themselves (the hypotheses about the evidence) often is called the prosecutor's fallacy,¹⁹² although instances are hardly confined to prosecutors.¹⁹³ Statisticians know it as the fallacy of the transposed conditional.¹⁹⁴ Despite the attention the transposition fallacy has received, no federal court has excluded a random-match probability as unfairly prejudicial simply because the jury might misinterpret it as a probability that the defendant is the

191. *McDaniel v. Brown*, 558 U.S. 120, 128 (2010) (per curiam).

192. *Id.* In *Brown*, a DNA analyst, at the prompting of the prosecution, suggested that a random-match probability of 1/3,000,000 implied a 0.000033 probability that the defendant was not the source of the DNA found on the victim's clothing. The Court unanimously held that a federal writ of habeas corpus should not have been issued in light of the limitations on federal review of state convictions. The prisoner argued, inter alia, that the transposed conditional probability made the DNA evidence so unreliable that its use deprived him of due process. The Court did not reach this issue, as it had not been raised below. The probabilities associated with the DNA evidence in the case are discussed in detail in Kaye, *Case Study*, *supra* note 162.

193. It also appears in statements from defense counsel, scientists, journalists, and judges. *E.g.*, *Williams v. Bauman*, 759 F.3d 630, 632 (6th Cir. 2014) ("a 1-in-7.663 quadrillion chance that it came from someone else"); *United States v. Ford*, 683 F.3d 761, 768 (7th Cir. 2012) ("[t]he probability that the DNA was someone else's would be 7 percent if the comparison were confined to the first location"); see also David H. Kaye et al., *The New Wigmore, A Treatise on Evidence: Expert Evidence* § 14.1.2(a) (2d ed. 2011) (chapter not included in the current edition) (collecting opinions expressing the fallacy); 2 Paul C. Giannelli et al., *Scientific Evidence* § 18.04[3][b][4], at 1893 n.384 ("nearly every appellate decision in the U.S. on the sufficiency of a DNA cold hit has fallen victim to the fallacy").

There is an opposing "defendant's fallacy" of dismissing or undervaluing matches because other matches are to be expected in unrealistically large populations of potential suspects. For example, defense counsel might argue that (1) with a random-match probability of one in a million, we would expect to find three or four unrelated people with the requisite genotypes in a major metropolitan area with a population of 3.6 million; (2) the defendant just happens to be one of these three or four, which means that the chances are at least 2 out of 3 that someone unrelated to the defendant is the source; so (3) the DNA evidence does nothing to incriminate the defendant. The problem with this argument is that in a case involving both DNA and non-DNA evidence against the defendant, it is unrealistic to assume that there are 3.6 million equally likely suspects. When juries are confronted with both fallacies, the defendant's fallacy seems to dominate. NRC II, *supra* note 7, at 198; cf. Jonathan J. Koehler, *The Psychology of Numbers in the Courtroom: How to Make DNA-Match Statistics Seem Impressive or Insufficient*, 74 S. Cal. L. Rev. 1275 (2001) (discussing ways of framing the evidence that make it more or less persuasive).

194. See Kaye & Stern, *supra* note 12, app'x section C (describing the fallacy and the relationship between the conditional probabilities); Ian W. Evett, *Avoiding the Transposed Conditional*, 35 Sci. & Just. 127 (1995).

source of the forensic DNA.¹⁹⁵ Courts, however, have noted the need to have the concept “properly explained,”¹⁹⁶ and prosecutorial or expert misrepresentations of the random-match probabilities for DNA and other trace evidence have produced reversals or contributed to the setting aside of verdicts.¹⁹⁷

Are Random-Match Probabilities That Are Smaller Than False-Positive Error Probabilities Irrelevant or Prejudicial?

Some scientists and lawyers have maintained that match probabilities are logically irrelevant when they are far smaller than the probability of a frame-up, a blunder in labeling samples, cross-contamination, or other events that would

195. See, e.g., *United States v. McCluskey*, 954 F. Supp. 2d 1224, 1291 (D.N.M. 2013); *United States v. Morrow*, 374 F. Supp. 2d 51, 66 (D.D.C. 2005) (“careful oversight by the district court and proper explanation can easily thwart this issue”).

196. *United States v. Shea*, 957 F. Supp. 331, 345 (D.N.H. 1997); see also *United States v. Chischilly*, 30 F.3d 1144, 1158 (9th Cir. 1994) (stating that the government must be “careful to frame the DNA profiling statistics presented at trial as the probability of a random match, not the probability of the defendant’s innocence that is the crux of the prosecutor’s fallacy”). The 1996 NRC committee suggested that “if the initial presentation of the probability figure, cross-examination, and opposing testimony all fail to clarify the point, the judge can counter [the fallacy] by appropriate instructions to the jurors that minimize the possibility of cognitive errors.” NRC II, *supra* note 7, at 198 (footnote omitted). The committee formulated the following instruction to define the random-match probability:

In evaluating the expert testimony on the DNA evidence, you were presented with a number indicating the probability that another individual drawn at random from the [specify] population would coincidentally have the same DNA profile as the [bloodstain, semen stain, etc.]. That number, which assumes that no sample mishandling or laboratory error occurred, indicates how distinctive the DNA profile is. It does not by itself tell you the probability that the defendant is innocent.

Id. at 198 n.93. An alternative adopted in England is to confine the prosecution to stating a frequency rather than a probability. See Kaye et al., *supra* note 193, § 14.1.2(b); cf. D.H. Kaye, *The Admissibility of “Probability Evidence” in Criminal Trials—Part II*, 27 *Jurimetrics J.* 160, 168 (1987) (similar proposal).

197. E.g., *United States v. Massey*, 594 F.2d 676, 681 (8th Cir. 1979) (explaining that in closing argument about hair evidence, “the prosecutor ‘confuse[d] the probability of concurrence of the identifying marks with the probability of mistaken identification’”) (alteration in original); cf. *Whack v. State*, 73 A.3d 186, 198 (Md. 2013) (reversing defendant’s conviction because of the prosecution’s closing argument about two probabilities and admonishing that “counsel have a responsibility to take extra care in describing DNA evidence, particularly when it comes to statistical probabilities”); *State v. Phillips*, 844 S.E.2d 651, 658 (S.C. 2020) (reversing partly as a result of inadequacies and errors in the presentation of random-match probabilities for “touch” DNA evidence and noting that “even when the concepts of touch DNA, non-exclusion DNA, and random match probability are completely and accurately presented to a jury, there is significant potential the testimony will be confusing and misleading”).

yield a false positive.¹⁹⁸ The argument is that the jury should concern itself only with the chance that the forensic sample is reported to match the defendant's profile even though the defendant is not the source. Match probabilities address only one path to such a false-positive report—a coincidence in correctly ascertained profiles. They do not consider the probability of a false-positive report because of fraud or an error in the collection, handling, or analysis of the DNA samples. Unless those probabilities are essentially zero, the match probability understates the chance of a reported match arising when DNA from a nonsource is compared to DNA from the source of the trace-evidence sample.

This mathematical observation has led to arguments that because probability of these other possible explanations for a match are larger than the very small random-match probabilities for most STR profiles, the latter probabilities are irrelevant. Commentators have crafted theoretical, doctrinal, and practical rejoinders to this claim.¹⁹⁹ The essence of the counterargument is that it is logical to give jurors information about kinship or random-match probabilities because, even if these numbers do not give the whole picture, they address pertinent hypotheses about the true source of the trace evidence.

It also has been argued that even if very small match probabilities are logically relevant, they are unfairly prejudicial in that they will cause jurors to neglect the probability of a match arising because of a false-positive laboratory error.²⁰⁰ A court that shares this concern might require the expert who presents a random-match probability also to report a probability that the laboratory is mistaken about

198. E.g., Jonathan J. Koehler et al., *The Random Match Probability in DNA Evidence: Irrelevant and Prejudicial?*, 35 *Jurimetrics J.* 201 (1995); Richard C. Lewontin & Daniel L. Hartl, *Population Genetics in Forensic DNA Typing*, 254 *Science* 1745, 1749 (1991), <https://doi.org/10.1126/science.1845040> (“[p]robability estimates like 1 in 738,000,000,000,000 . . . are terribly misleading because the rate of laboratory error is not taken into account”).

199. See Kaye et al., *supra* note 193, § 14.1.1 (discussing the issue).

200. Some commentators believe that this prejudice is so likely and so serious that “jurors ordinarily should receive *only* the laboratory’s false positive rate. . . .” Richard Lempert, *Some Caveats Concerning DNA as Criminal Identification Evidence: With Thanks to the Reverend Bayes*, 13 *Cardozo L. Rev.* 303, 325 (1991) (emphasis added). The 1996 NRC committee was skeptical of this view, especially when the defendant has had a meaningful opportunity to retest the DNA at a laboratory of his or her choice, and it suggested that judicial instructions can be crafted to avoid this form of prejudice. NRC II, *supra* note 7, at 199. Pertinent psychological research includes Dale A. Nance & Scott B. Morris, *Juror Understanding of DNA Evidence: An Empirical Assessment of Presentation Formats for Trace Evidence with a Relatively Small Random Match Probability*, 34 *J. Legal Stud.* 395 (2005); Dale A. Nance & Scott B. Morris, *An Empirical Assessment of Presentation Formats for Trace Evidence with a Relatively Large and Quantifiable Random Match Probability*, 42 *Jurimetrics J.* 403 (2002); Jason Schklar & Shari Seidman Diamond, *Juror Reactions to DNA Evidence: Errors and Expectancies*, 23 *Law & Hum. Behav.* 159, 179 (1999), <https://doi.org/10.1023/A:1022368801333> (concluding that separate figures for laboratory error and a random match to a correctly ascertained profile are desirable in that “[j]urors . . . may need to know the disaggregated elements that influence the aggregated estimate as well as how they were combined in order to evaluate the DNA test results in the context of their background beliefs and the other evidence introduced at trial”).

the profiles. Of course, for reasons given in the section titled “How Are Samples Created and Handled?” above, some experts would deny that they can provide a meaningful statistic for the case at hand, but they could report the results of proficiency tests and leave it to the jury to use this figure as best it can in considering whether a false-positive error has occurred.²⁰¹ In any event, the courts have been unreceptive to efforts to replace random-match probabilities with a blended figure that incorporates the risk of a false-positive error²⁰² or to exclude random-match probabilities that are not accompanied by a separate false-positive error probability.²⁰³

“Rarity,” Source, and Uniqueness Testimony

Having surveyed the issues related to the value and dangers of probabilities and statistics for DNA evidence, we turn to a related issue that can arise under Rules 702 and 403: Should an expert be permitted to offer a nonnumerical judgment about the DNA profiles? Many courts have held that a DNA match is inadmissible unless the expert attaches a scientifically valid number to the match. Indeed, some opinions state that this requirement flows from the nature of science

201. *Cf.* *Williams v. State*, 679 A.2d 1106, 1120 (Md. 1996) (reversing because the trial court restricted cross-examination about the results of proficiency tests involving other DNA analysts at the same laboratory). *But see* *United States v. Shea*, 957 F. Supp. 331, 344 n.42 (D.N.H. 1997) (“The parties assume that error rate information is admissible at trial. This assumption may well be incorrect. Even though a laboratory or industry error rate may be logically relevant, a strong argument can be made that such evidence is barred by Fed. R. Evid. 404 because it is inadmissible propensity evidence.”).

202. *United States v. McCluskey*, 954 F. Supp. 2d 1224, 1270–73 (D.N.M. 2013); *United States v. Ewell*, 252 F. Supp. 2d 104, 113–14 (D.N.J. 2003); *United States v. Shea*, 957 F. Supp. 331, 334–45 (D.N.H. 1997); *People v. Reeves*, 109 Cal. Rptr. 2d 728, 753 (Ct. App. 2001); *State v. Tester*, 968 A.2d 895 (Vt. 2009); *cf.* *Armstead v. State*, 673 A.2d 221, 245 (Md. 1996) (the failure to combine a random-match probability with an error rate on proficiency tests that was many orders of magnitude greater (and that was placed before the jury) did not deprive the defendant of due process). Formulas for incorporating error into the likelihood are given in Tacha Hicks et al., *A Framework for Interpreting Evidence*, in *Forensic DNA Evidence Interpretation* 37, 73 (John Buckleton et al. eds., 2d ed. 2016), <https://doi.org/10.1201/b19680-3>.

203. *McCluskey*, 954 F. Supp. 2d at 1270–73; *United States v. Trala*, 162 F. Supp. 2d 336, 350–51 (D. Del. 2001); *United States v. Lowe*, 954 F. Supp. 401, 415–16 (D. Mass. 1997), *aff’d*, 145 F.3d 45 (1st Cir. 1998) (a “theoretical” error rate need not be presented when quality-assurance standards have been followed and defendant had the opportunity to retest the sample); *Roberts v. United States*, 916 A.2d 922, 930–31 (D.C. 2007); *Roberson v. State*, 16 S.W.3d 156, 168 (Tex. Crim. App. 2000) (error rate not needed when laboratory was accredited and underwent blind proficiency testing); *Tester*, 968 A.2d 895 (finding when the laboratory chemist stated that “[t]here is no error rate to report” because the number of proficiency trials was insufficient, the random-match probability was admissible and preferable to presenting the finding of a match with no accompanying statistic).

itself.²⁰⁴ However, this view has been challenged,²⁰⁵ and not all courts agree that an expert must explain the power of a DNA match in purely numerical terms.²⁰⁶

Instead of presenting numerical frequencies or match probabilities, a scientist could characterize a multilocus STR profile as “rare,” “extremely rare,” or the like.²⁰⁷ A few opinions suggest that a purely qualitative presentation would be admissible,²⁰⁸ but with the availability of statistically sound estimates of frequencies or conditional probabilities, such testimony is itself quite rare. When numerical estimates are possible, it is not clear what the expert’s verbal tag accomplishes.

The most extreme case of a purely verbal description of the infrequency of a profile occurs when that profile can be said to be unique. Of course, the uniqueness of any object, from a snowflake to a fingerprint, in a population that cannot be enumerated, never can be proved directly. As with all sample evidence, one must generalize from the sample to the entire population. There is always some probability that a census would prove the generalization to be false. Almost three decades ago, the second National Research Council (NRC) committee therefore wrote:

[T]here is no “bright-line” standard in law or science that can pick out exactly how small the probability of the existence of a given profile in more than one member of a population must be before assertions of uniqueness are justified. . . . There might already be cases in which it is defensible for an expert to assert that, assuming that there has been no sample mishandling or laboratory error, the profile’s probable uniqueness means that the two DNA samples come from the same person.²⁰⁹

204. *E.g.*, *State v. Cauthron*, 846 P.2d 502 (Wash. 1993).

205. *See, e.g.*, *State v. Hummert*, 933 P.2d 1187, 1191 (1997) (“We believe this view to be seriously incorrect.”); *Commonwealth v. Crews*, 640 A.2d 395, 402 (Pa. 1994) (“[T]he factual evidence of the physical testing of the DNA samples and the matching alleles, even without statistical conclusions, tended to make appellant’s presence more likely than it would have been without the evidence, and was therefore relevant.”). The National Research Council committee wrote that science only demands “underlying data that permit some reasonable estimate of how rare the matching characteristics actually are,” and “[o]nce science has established that a methodology has some individualizing power, the legal system must determine whether and how best to import that technology into the trial process.” NRC II, *supra* note 7, at 192.

206. *E.g.*, *Rodriguez v. State*, 273 P.3d 845, 851 (Nev. 2012); *People v. Her*, 157 Cal. Rptr. 3d 40 (Cal. Ct. App. 2013). For discussion of pure “defendant-not-excluded” testimony, see *United States v. Morrow*, 374 F. Supp. 2d 51 (D.D.C. 2005); *Kaye et al.*, *supra* note 193, § 15.4.

207. *Banks v. State*, 219 So.3d 19, 28 (Fla. 2017) (“by the time you get to 13 [STR loci] it becomes extremely rare, extremely”); *cf.* *State v. Hummert*, 933 P.2d 1187, 1193 (Ariz. 1997) (testimony that “of their own experience, they believed such a [three-locus RFLP-VNTR] random match would be very uncommon” was admissible).

208. *State v. Bloom*, 516 N.W.2d 159, 166–67 (Minn. 1994) (“Since it may be pointless to expect ever to reach a consensus on how to estimate, with any degree of precision, the probability of a random match and that, given the great difficulty in educating the jury as to precisely what that figure means and does not mean, it might make sense to simply try to arrive at a fair way of explaining the significance of the match in a verbal, qualitative, nonquantitative, nonstatistical way.”).

209. NRC II, *supra* note 7, at 194.

Before concluding that a DNA profile is unique in a given population, however, a careful expert also should consider not only the random-match probability (which pertains to unrelated individuals) but also the chance of a match to a close relative. Indeed, the possible existence of an unknown, identical twin also means that a scientist never can be absolutely certain that crime-scene evidence could have come from only the defendant.

Courts have accepted or approved of expert assertions of uniqueness or of individual-source identification.²¹⁰ For these assertions to be justified, a large number of sufficiently polymorphic loci must have been tested, making the probabilities of matches to both relatives and unrelated individuals so tiny that the probability of finding another person who could be the source within the relevant population is negligible.²¹¹ But deciding what probability is negligible is a personal or policy judgment rather than an objective fact, and trusting the probability models at the extremes where they cannot be experimentally tested

210. *E.g.*, United States v. Carney, No. 1:20-cr-121, 2022 WL 1155902, at *1 (S.D. Ohio Apr. 19, 2022) (“in the absence of an identical twin, Furious Carney is the source of the major DNA profile”); United States v. Davis, 602 F. Supp. 2d 658 (D. Md. 2009) (“the random match probability figures . . . are sufficiently low so that the profile can be considered unique”); People v. Cordova, 358 P.3d 518, 538 (2015) (because “the odds were astronomical,” an “opinion that defendant was the source of the evidentiary samples to a reasonable scientific certainty was reasonable”); State v. Hauge, 79 P.3d 131 (Haw. 2003) (uniqueness); Young v. State, 879 A.2d 44, 46 (Md. 2005) (holding that “when a DNA method analyzes genetic markers at sufficient locations to arrive at an infinitesimal random match probability, expert opinion testimony of a match and of the source of the DNA evidence is admissible”; hence, it was permissible to introduce a report providing no statistics but stating that “(in the absence of an identical twin), Anthony Young (K1) is the source of the DNA obtained from the sperm fraction of the Anal Swab (R1)”); State v. Buckner, 941 P.2d 667, 668 (Wash. 1997) (in light of 1996 NRC Report, “we now conclude there should be no bar to an expert giving his or her expert opinion that, based upon an exceedingly small probability of a defendant’s DNA profile matching that of another in a random human population, the profile is unique”).

211. Three distinct probabilities arise in speaking of the uniqueness of DNA profiles. First, there is the probability of a match to a single, randomly selected individual in the population. This is the random-match probability. Second, there is the probability that the particular profile is unique. This probability involves pairing the profile with every member of the population. Third, there is the probability that all pairs of all profiles are unique. The first probability is larger than the second, which is many times larger than the third. Uniqueness or source testimony need only establish that *the one* DNA profile in the trace evidence is unique—and not that *all* DNA profiles are unique. Thus, it is the second probability, properly computed, that must be quite small to warrant the conclusion that no one but the defendant (and any identical twins) could be the source of the crime-scene DNA. See David H. Kaye, *Identification, Individuality, and Uniqueness: What’s the Difference?*, 8 Law, Probability & Risk 85 (2009), <https://doi.org/10.1093/lpr/mgp018>.

Formulas for estimating all these probabilities are given in NRC II, *supra* note 7, but DNA analysts and judges sometimes infer uniqueness on the basis of incorrect intuitions about the size of the random-match probability. See David J. Balding, *Weight-of-Evidence for Forensic DNA Profiles* 148 (2005) (describing “the uniqueness fallacy”); cf. State v. Lee, 976 So. 2d 109, 117 (La. 2008) (incorrect but harmless miscalculation).

makes some scientists wary of claiming uniqueness and making source attributions on the basis of DNA evidence.²¹²

Special Issues in Human DNA Testing

Y Chromosomes

Genetic Principles

To understand how Y chromosomes are used in forensic-DNA science, we need to recall a few principles of the genetics of sexual reproduction mentioned in the section above titled “Sexual Reproduction and the Genome.” A basic point is that a male child receives an X chromosome from the genetic mother and a Y from the genetic father,²¹³ whereas female children receive two X chromosomes, one from each parent.²¹⁴ Thus, genetic males are type XY and genetic females are XX.²¹⁵

A second fundamental point is that every sex cell (an egg or a sperm) has only one copy of each parental chromosome, selected at random from the pairs carried by each parent. Hence, about half the sperm cells in a genetic male have a Y chromosome, whereas the male’s bodily (somatic) cells all have the full XY pair of chromosomes.

Third—and this is crucial in forensic identification with Y chromosomes—when a sperm cell forms, there is limited recombination (exchanges of segments) between X and Y chromosomes. Along most of its length, the Y chromosome stays intact, and only occasional intergenerational mutations produce diversity among Y chromosomes within a population. In other words, the Y loci used in forensic identification are linked rather than independent. They are inherited as a single block—a haplotype—from father to son, and all the men in the same paternal line (up to the last mutation giving rise to a new line in the family tree) would match the trace-evidence sample’s Y haplotype. At the same time, men from a different paternal lineage (one with no recent male ancestor in common) might have developed the same haplotype (depending on the history of mutations

212. John S. Buckleton et al., *Single Source Samples, in* Forensic DNA Evidence Interpretation 203, 207–15 (John S. Buckleton et al. eds., 2d ed. 2016).

213. All told, the Y chromosome represents about 2% of the total human genome and is much smaller than its X counterpart.

214. This pattern is a consequence of the fact that certain genes in the Y chromosome result in development as a male rather than a female.

215. Sometimes, progeny are born with fewer or more than the usual two sex chromosomes. Jeannie Visootsak & John M. Graham, Jr., *Klinefelter Syndrome and Other Sex Chromosomal Aneuploidies*, 1 Orphanet J. Rare Diseases 42 (2006), <https://doi.org/10.1186/1750-1172-1-42>; Brittany Sood & Roselyn W. Clemente Fuentes, *Jacobs Syndrome*, StatPearls (2022), <https://perma.cc/62WT-W4S3>; *supra* note 22.

giving rise to each line). If so, men from both lines would match a trace-evidence sample with the shared haplotype.

Like the other 22 paired chromosomes (the autosomes), the Y chromosome contains not only genes but also STRs, SNPs, and other parts that vary from one person (or group of people) to the next. A particular haplotype is thus some combination of the variants at some of these loci.

Forensic Value

In directly comparing DNA profiles, there normally is not much point in adding Y-STRs to the autosomal profile. The rarity of the multilocus autosomal STR genotypes already makes them very discriminating.²¹⁶ Nevertheless, the patrilineal inheritance of Y chromosomes can be useful for determining the number of male contributors to a mixed semen sample in sexual assault cases; for discerning which male haplotype is present when a genetic male's autosomal STR alleles cannot be detected in the midst of a much larger quantity of female DNA;²¹⁷ for familial searching in criminal DNA databases;²¹⁸ for inferring biogeographical ancestry as an investigative lead;²¹⁹ and for establishing kinship in immigration cases or other situations.²²⁰ A famous example of the last application is the Y-chromosome testing of members of the acknowledged families of President Thomas Jefferson and Sally Hemings that revealed Sally's last child had a fairly rare Y haplotype that President Jefferson—and everyone else in the Jefferson male clan—possessed.²²¹

216. Moreover, some geneticists are not confident that Y-STRs are independent of autosomal STRs, making it unclear how to arrive at an estimate for a profile frequency or probability. See Weir, *supra* note 182, at 258 (“[A]lthough the dependencies [among autosomal STRs and Y-chromosome and mitochondrial systems] are low, they do exist. Combining match probability estimates over systems is not generally recommended.”).

217. *E.g.*, State v. Jones, 345 P.3d 1195 (Utah 2015) (DNA on fingernail clippings from female homicide victim who struggled with her attacker); State v. Maestas, 299 P.3d 892 (Utah 2012) (same).

218. See *infra* section titled “Kinship trawling.”

219. See *infra* section titled “Biogeographic ancestry testing.”

220. For example, to help identify human remains as those of a male reported as missing, DNA from the remains might be compared to a sample from a father, brother, uncle, or other relative in the same paternal line as the missing person. The presence of the same Y-STRs in both samples tends to show that the remains are indeed those of the missing person.

221. Eugene A. Foster et al., *Jefferson Fathered Slave's Last Child*, 396 Nature 27 (1998), <https://doi.org/10.1038/23835>. In response to letters emphasizing the inability of markers on the Y chromosome to distinguish between members of the Jefferson clan (including any illegitimate ones, white or black), the authors of the study noted that the genetic information was much more probable if the President rather than one of his sister's children fathered Sally's last child. The paper's title, they acknowledged, was misleading. Eugene A. Foster et al., *The Thomas Jefferson Paternity*

Analytical Methods and Categorical Matches

Typically, the loci used in constructing a Y haplotype for forensic applications are STRs. Up to 30 are in use. For capillary electrophoresis, the same instruments and software used for routine STR testing are employed. Large peaks in an electropherogram that do not seem to be artifacts are interpreted as the alleles of the haplotype.²²² As with autosomal STR profiling, the common practice for comparing a suspect's profile to a trace-evidence one is the two-stage framework of a categorical match or no-match conclusion followed by a statistical evaluation. For instance, the Department of Justice's Uniform Language for Testimony and Reports (ULTR) for Y-STRs states that an examiner "may" testify to an inclusion (meaning "included, or cannot be excluded"), exclusion, or inconclusive (because of an "indistinguishable mixture" or a possible "mutational event").²²³ A declared match must be accompanied by some kind of "quantitative statement describing the weight of the evidence."²²⁴

Case, 397 Nature 32 (1999), <https://doi.org/10.1038/16177>; see also Eliot Marshall, *Which Jefferson Was the Father?*, 283 Science 153 (1999), <https://doi.org/10.1126/science.283.5399.153a>.

222. Analysts need to be aware of certain quirks with Y-STR profiling to avoid some possible misinterpretations of electropherograms. The International Society of Forensic Genetics explains that

[S]ome mutational effects rarely seen in autosomal STRs are more pronounced in Y-STRs. Especially large-scale deletions, insertions and conversions are responsible for higher numbers of Null . . . and multiple alleles at certain loci. Some markers included in commercial kits show always more than one allele because the sequence has identical copies on the Y chromosome (e.g. DYS385, DYS387S1). One or several Y-STR loci per haplotype can be involved in such mutation events. This is of forensic relevance, because a pattern can be erroneously interpreted as allelic drop-out, DNA contamination[,] and mixture, which may affect the evidential value of a DNA profile.

Lutz Roewer et al., *DNA Comm'n of the Int'l Soc'y of Forensic Genetics (ISFG): Recommendations on the Interpretation of Y-STR Results in Forensic Analysis*, 48 Forensic Sci. Int'l: Genetics 102308, at 2 (2020), <https://doi.org/10.1016/j.fsigen.2020.102308> (citations omitted). SWGDAM grandly declares that "[t]he laboratory should establish a method based on validation to document the designation of a peak as an artifact or an allele." Scientific Working Group on DNA Analysis Methods, Interpretation Guidelines for Y-Chromosome STR Typing by Forensic DNA Laboratories § 4.1 (Mar. 2, 2022), <https://perma.cc/ZZR4-FPND>. It discusses the difficulties further in related and supplemental documents.

223. U.S. Dep't of Just., Uniform Language for Testimony and Reports for Forensic Y-STR DNA Examinations 2–3 (Dec. 12, 2022) (<https://perma.cc/LE6X-FS87>) [hereinafter Y ULTR].

224. *Id.* at 3 ("An examiner shall provide a quantitative statement describing the weight of the evidence for all comparisons in which a known male is included as a possible contributor to the Y-STR typing results obtained from a probative evidentiary sample. This statement shall be provided regardless of the number of alleles detected or the magnitude of the resulting quantitative value."). The ISFG is more lenient. Roewer et al., *supra* note 222, at 2 ("Basically, examiners are expected to prepare reports with the minimum requirement of a *qualitative statement* on the Y-STR test result."). But see Y ULTR, *supra* note 223, at 5 ("When a suspect is identified that matches the Y-STR profile, we recommend the formulation of hypotheses according to the likelihood approach described by Evett and Weir."); see also Ian W. Evett & Bruce S. Weir, *Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists* (1998).

In court, laboratory performance issues and the manner in which inclusionary findings are presented are more likely to be raised than are any fundamental objections to the science and technology for determining whether two Y-STR profiles are the same. Reasoning that Y-STRs are just another set of markers involving the same principles and procedures as the previously accepted autosomal STRs, courts have upheld the admission of matching Y-STRs.²²⁵ Likewise, actions for postconviction Y-STR testing tend to be resolved on the assumption that this form of DNA testing is admissible.²²⁶

Population Genetics and Statistics

An exclusion reflects the analyst's belief that differences between the alleles in the profiles are so extensive that they would never be seen in samples within the same paternal lineage. The inference that follows is clear: A correct exclusion means the suspect is not the source. However, the implication of an inclusion or match is not obvious without further information. If the inclusion is correct, the haplotypes are the same. But what follows from that? Does the shared haplotype occur once in a blue moon in the suspect population, or does it pop up all the time? Thus, an inclusion requires a second stage of statistical analysis to know the degree to which the finding, even if correct, tends to prove that the defendant is the source of the trace-evidence sample.²²⁷ The additional information

225. *E.g.*, *Shabazz v. State*, 592 S.E.2d 876, 879 (Ga. Ct. App. 2004); *Commonwealth v. Jacoby*, 170 A.3d 1065, 1091–95 (Pa. 2017) (nothing novel about Y-STRs); *Curtis v. State*, 205 S.W.3d 656, 660–61 (Tex. Ct. App. 2006); *Commonwealth v. Clark*, 34 N.E.3d 1, 12 (Mass. 2015) (stating “that the results of DNA testing using the Y-STR method are admissible in Massachusetts courts” on the basis of a case with autosomal-STR testing).

226. *E.g.*, *United States v. MacDonald*, 37 F. Supp. 3d 782 (E.D.N.C. 2014) (motion for Y-STR typing under the Innocence Protection Act of 2004, 18 U.S.C. § 3600, denied as untimely); *Clark*, 34 N.E.3d 1 (denial of a motion for Y-STR testing under a state statute was error).

227. An “inconclusive” means that the DNA test should not be used to decide whether the suspect is the source. It reflects the belief that even if the samples are in the same lineage, the observed differences are not extremely improbable considering the possibility of a mutation or other event. Neither are they very improbable if the samples come from different paternal lines. No inference follows. Yet, DNA analysts sometimes speak of the proportion of a population that would be excluded by haplotype testing as if it were simply 1 minus the proportion of fully matching haplotypes in a reference database for that population. *E.g.*, *State v. Polizzi*, 924 So.2d 303, 308–09 (La. Ct. App. 2006) (The DNA expert “concluded that the two profiles were consistent with each other, meaning that the Defendant or any of his paternal relatives could not be excluded as having been a donor to the sample from the victim. . . . 99.7 percent of the Caucasian population, 99.8 percent of the African American population, and 99.3 percent of the Hispanic population could be excluded as donors of the DNA in the sample.”). But if haplotypes with slightly different STRs—they match except for one mutation—are considered inconclusive, men with those haplotypes cannot be excluded. Hence, the one-off haplotypes should be added in computing the fraction of the population that “could not be excluded.” Whether such slightly different haplotypes

needed to evaluate a match comes from population genetics and statistics. To provide this information, researchers have gathered DNA samples from populations both within the United States and across the world and examined them for Y-STRs and Y-SNPs.²²⁸ The goal is to draw on these “reference databases” to estimate the unrelated-man-haplotype frequency or match probability.

Haplotype frequency or probability estimates

Because the entire haplotype (for example, a set of perhaps a dozen to 30 STRs strung out across the Y chromosome) is inherited as a package, it is neither necessary nor appropriate to combine any allele frequencies as is done with autosomal STRs. The haplotype itself is like a single allele, and one can simply count how often it occurs in a sample of unrelated men.²²⁹ As discussed in the *Reference Guide on Statistics and Research Methods*, in this manual, dividing the count (x) by the sample size (n) gives the sample proportion ($p = x/n$). For instance, in *State v. Jones*,²³⁰ a Y-STR haplotype was present in DNA from fingernail clippings from a woman who had been murdered in the Salt Lake City area and who evidently had struggled with her attacker. A laboratory director testified to zero occurrences of the haplotype in a database of size $n = 8,028$.²³¹ The sample proportion in *Jones* was therefore $p = 0/8,028 = 0$.

Two variations on the sample proportion x/n as an estimator of a population frequency have some popularity in the forensic-genetics community. Following the testing in the case, the number of times that the haplotype has been seen is no longer x out of n . Now it is $x + 1$ out of the $n + 1$ men with known haplotypes. In *Jones*, this number is $1/8,029$, or about 0.012%. Internationally, this “augmented” statistic is typically used as an estimator of the population frequency.²³² There also are arguments for using $(x + 2)/(n + 2)$ or $(x + 2)/(n + 3)$, which would be about 0.025% in *Jones*.²³³

were treated as “consistent with” one another in the percentages recited in *Polizzi* is ambiguous at best.

228. The freely available Y Chromosome Haplotype Reference Database contains data from “more than 1,300 local population samples with more than 270,000 haplotypes in 136 countries.” Lutz Roewer, *Using the YHRD Database for Casework Analysis*, ISHI Rep. Apr. 2019, <https://perma.cc/25AP-F6VJ>.

229. In this context, “unrelated people” means individuals with a different paternal lineage.

230. 345 P.3d 1195 (Utah 2015).

231. *Id.* at 1208 n.42.

232. Roewer et al., *supra* note 222, at 3.

233. See Michael Coble et al., *Non-autosomal Forensic Markers*, in *Forensic DNA Evidence Interpretation* 315, 331 (John Buckleton et al. eds., 2d ed. 2016). The SWGDAM Y chromosome guidelines only state that “[t]he laboratory should determine which of the following methods will be used.” SWGDAM, *supra* note 222, § 9.2.2.

These sample proportions do not account for sampling error—the inevitable differences between random samples and the population from which they are drawn. Finding no matches in a sample of 8,028 men hardly means that there would be no other paternal line with a matching haplotype among the hundreds of thousands of men in the Salt Lake City area. If there are any, a different sample might have produced some of them. One solution to this “zero numerator” problem²³⁴ (and to the problem of quantifying sampling error generally) is to supply a confidence interval for plausible values of the population frequency. Testimony that incorporates this concept is common. The laboratory director in *Jones* testified “that ‘approximately 99.6 percent of . . . the male population can be excluded’ as a contributor of the DNA sample but that Mr. Jones could not be excluded” and that “read another way, the frequency of Mr. Jones’s DNA profile ‘is equivalent to one in 2681 individuals.’ He explained this means that ‘every time you test . . . a male [in a different paternal line], the probability of that person having that particular DNA profile is approximately one in 2681.’”²³⁵

How did the witness get from about 0/8,000 to 1/2,681? The latter figure is an answer to the following question: How large would the population proportion have to be before we would expect to encounter random samples of size 8,028 that are devoid of the haplotype (as in our one database of this size) at least 5% of the time? If the proportion were much larger than 1/2,681, then random samples with no matching haplotype would be rarer. As such, an estimate of 1/2,681 for the population haplotype frequency is not plainly incompatible with the finding that the haplotype is not in the database. It is the upper end of a 95% “confidence interval.”²³⁶

Case law generally accepts the confidence-interval procedure for estimating the frequency of a Y haplotype in reference-population databases.²³⁷ It also rejects arguments that quantitative analyses are unfairly prejudicial; however,

234. Kaye & Stern, *supra* note 12, section titled “Other Situations.”

235. *State v. Jones*, 345 P.3d 1195, 1208 (Utah 2015) (omissions in original).

236. When drawing from a population whose proportion is 1/2,681, the probability of a simple random sample of size 8,028 with no matching haplotypes is 5%. For a population proportion of 1/1,744, the probability of no haplotypes is 1%, so it can be called the upper limit of a 99% confidence interval. A 99% interval may sound better than a 95% interval, but the larger proportion of 1/1,744 is less compatible with the database. It generates random samples that lack the haplotype less often.

The calculations here use the Poisson approximation. Other ways to produce confidence intervals are sensible. Coble et al., *supra* note 233, at 338–39; Kaye & Stern, *supra* note 12, section titled “Other Situations.” The SWGDAM guidelines cite one procedure. SWGDAM, *supra* note 222, § 9.2.3.

237. *E.g.*, *Commonwealth v. Jacoby*, 170 A.3d 1065 (Pa. 2017); *State v. Tucker*, 920 N.W.2d 680 (Neb. 2018); *State v. Jones*, 345 P.3d 1195 (Utah 2015); *State v. Bander*, 208 P.3d 1242, 1255 (Wash. Ct. App. 2009) (noting that defendant made no showing of a scientific controversy); *cf.* *Commonwealth v. Lally*, 46 N.E.3d 41, 52–53 (Mass. 2016) (confidence interval not required if point estimate is explained clearly).

some opinions admonish trial judges and counsel to avoid exaggerating the power of Y haplotypes for source attribution.²³⁸ The scientific consensus seems to be that the “counting method” and its confidence intervals are overly conservative for rare haplotypes.²³⁹ As such, further refinements have been developed,²⁴⁰ and population-genetics models that take into account the way haplotypes evolve in reproducing populations may be the most technically defensible basis for estimating frequencies and likelihood ratios.²⁴¹

Likelihood ratios (LRs)

Likelihood ratios (defined in the section titled “Mixtures” later in this reference guide) are recommended for Y haplotypes by the International Society of Forensic Genetics (ISFG),²⁴² are recognized by SWGDAM,²⁴³ and are advocated for reporting DNA and other forensic-science findings by many scholars and other groups.²⁴⁴ LRs have the potential advantage of clarifying the weight of the evidence with respect to a range of alternative hypotheses. For example, the contribution of the analysis of the Y haplotypes to the historical debate over the paternity of Sally Hemings’s children (see section titled “Forensic Value” above)

238. *Jones*, 345 P.3d at 1249.

239. Roewer et al., *supra* note 222, at 2–3; Coble et al., *supra* note 267, at 330–35.

240. One such adjustment is the “kappa method.” Butler, *supra* note 113, at 423–24; Coble et al., *supra* note 267, at 332; Roewer et al., *supra* note 222, at 3.

241. The discrete Laplace method has been said to be the most robust. Coble et al., *supra* note 267, at 333; *see also* Roewer et al., *supra* note 222, at 3 (“established in practice” internationally). Another consideration is the adjustment for population structure when the reference database comes from an amalgam of preferentially reproducing subgroups in a larger population (see *supra* section titled “Could an Unrelated Person Be the Source?”). For advice, see Coble et al., *supra* note 233, at 329, 330, 334–35; Roewer et al., *supra* note 222, at 5.

242. Roewer et al., *supra* note 222, at 5. The ISFG gives a generic example of possible wording to explain the number:

The Y-STR profile detected in the crime stain is LR times more probable to observe under hypothesis H_1 than under hypothesis H_2 . This notwithstanding, paternal relatives have a high probability to have the same Y-STR profile and will in that case have the same likelihood ratio (LR).

Id.

243. SWGDAM, *supra* note 222, § 9.2.5 (“The laboratory should determine if likelihood ratios will be used to provide quantitative assessments of the value of the matches using relevant populations.”).

244. *See* section titled “Frequencies, Probabilities, and Prejudice” above. The Department of Justice ULTR demands that “[a]n examiner shall provide a quantitative statement describing the weight of the evidence for all comparisons in which a known male is included as a possible contributor to the Y-STR typing results obtained from a probative evidentiary sample.” Y ULTR, *supra* note 223, at 3. Presumably, the LR can be the “quantitative statement [of] weight” for inclusions, but the ULTR does not address its use for exclusions (where the LR is very large for the hypothesis that a man other than the defendant is the source as opposed to the hypotheses that the defendant is the source) and inconclusives (where the LR is 1).

might have been clearer from the outset had the researchers made statements for each hypothesis about paternity—for example:

- The Jefferson haplotype is at least 100 times more probable if the president was the father of Eston Hemings Jefferson than if one of the president's sister's boys was the father;
- The Jefferson haplotype has the same probability if the president was the father as it does if the president's brother was the father.

Similarly, in *State v. Jones*,²⁴⁵ the likelihood approach to reporting results would transform the testimony that “this means that 99.92 percent of the male population could be excluded as a possible donor” into:

- The haplotype is about 2,680 times more probable if Mr. Jones is its source than if any particular man not in Mr. Jones's paternal family tree is the source; and
- The haplotype has the same probability if Mr. Jones is its source as it does if any other man in his paternal family tree (a father, child, uncle, certain cousins, nephews, etc.) is the source.

In addition to its clarity with regard to hypotheses in the case, the broader likelihood ratio framework does not require the match/no-match categories and rules for “inconclusives.” It can better describe the evidentiary value of haplotypes that are one mutation away from a clear match²⁴⁶ and can be applied to ambiguous data from low-template DNA.²⁴⁷ However, questions can remain as to the computation of likelihood ratios in these more challenging cases.

Mitochondrial DNA

Mitochondria and Their Genomes

Mitochondria are small structures, with their own membranes, found inside the cell but outside its nucleus. Within these organelles, molecules are broken down to supply energy. Mitochondria have a small genome—a circle of 16,569 nucleotide base pairs that includes only 37 genes. They started out as bacteria that were engulfed by cells billions of years ago. Many of their original genes migrated to

245. See *supra* section titled “Genetic Principles.”

246. Coble et al., *supra* note 233, at 340; cf. *supra* note 227 (on the impact of “inconclusives” on exclusion probabilities).

247. Coble et al., *supra* note 233, at 342–44.

the nuclear chromosomes,²⁴⁸ where they produce proteins and RNAs for the mitochondria.²⁴⁹ However, the sequences currently found in the mitochondria themselves bear little relation to the comparatively monstrous chromosomal genome in the cell nucleus. In particular, they are not physically or statistically associated with the forensic STRs.

Forensic Value

Mitochondrial DNA (mtDNA) has four features that make it useful for forensic DNA testing. First, the typical cell, which has but one nucleus, contains hundreds or thousands of nearly identical mitochondria. Hence, for every copy of chromosomal DNA, there are hundreds or thousands of copies of mitochondrial DNA. This makes it possible to detect mtDNA in samples, such as bone and hair shafts, that contain too little nuclear DNA for conventional typing.

Second, two hypervariable regions that tend to be different in different individuals lie within the control region or D-loop (displacement loop) of the mitochondrial genome.²⁵⁰ These regions extend for a bit more than 300 base pairs each—short enough to be typable even in highly degraded samples such as very old human remains.

Third, mtDNA comes solely from the egg cell.²⁵¹ For this reason, mtDNA is inherited maternally, with no paternal contribution.²⁵² Full siblings, maternal half-siblings, and others related through maternal lineage normally possess the same mtDNA sequence. This feature makes mtDNA particularly useful for

248. James B. Stewart & Patrick F. Chinnery, *Extreme Heterogeneity of Human Mitochondrial DNA from Organelles to Populations*, 22 *Nature Rev. Genetics* 106, 106–07 (2021), <https://doi.org/10.1038/s41576-020-00284-x>.

249. Whole-genome sequencing of parents and their children establishes that other segments of mitochondrial DNA (mtDNA) continue to find their way into the nuclear genome, creating new nuclear mtDNA segments (NUMTs) in some cells. *E.g.*, Wei Wei et al., *Nuclear-embedded Mitochondrial DNA Sequences in 66,083 Human Genomes*, 611 *Nature* 105 (2022), <https://doi.org/10.1038/s41586-022-05288-7>.

250. A third, somewhat less polymorphic region in the D-loop, can be used for additional discrimination. The remainder of the control region, although noncoding, consists of DNA sequences that are involved in the transcription of the mitochondrial genes. These control sequences are essentially the same in everyone (monomorphic).

251. The relatively few mitochondria in the spermatozoan that fertilizes the egg cell soon degrade and are not replicated in the multiplying cells of the pre-embryo.

252. The possibility of paternal contributions to mtDNA in humans is discussed in Alistair T. Pagnamenta et al., *Biparental Inheritance of Mitochondrial DNA Revisited*, 22 *Nature Rev. Genetics* 477 (2021), <https://doi.org/10.1038/s41576-021-00380-6> (“Evidence for a biparental mode of . . . inheritance has been sparse and remains controversial. Recent studies using a range of complementary techniques do not support paternal transmission of mtDNA and highlight the co-amplification of rare, concatenated nuclear mtDNA segments as a technical artefact that may explain previous observations.”).

associating persons related through their maternal lineage. It has been exploited to identify the remains of the last Russian tsar and other members of the royal family, of soldiers missing in action, and of victims of mass disasters.²⁵³

Finally, point mutations accumulate, particularly in the noncoding D-loop, without altering how the mitochondrion functions. Hence, a single individual can develop distinct internal populations of mitochondria. Single nucleotide and other variants can occur within a cell, between cells in a given tissue, and between cells from different tissues and organs.²⁵⁴ Such “heteroplasmy” has been observed in healthy humans, with an average of one heteroplasmic site per individual.²⁵⁵ As discussed below, heteroplasmy complicates the interpretation of differences in mtDNA sequences.²⁵⁶ Yet, it is mutations that make mtDNA polymorphic and hence useful in identifying individuals. Over time, mutations in egg cells can propagate to later generations, producing more heterogeneity in the inherited mitochondrial genomes in the human population.²⁵⁷ This polymorphism allows scientists to compare mtDNA from crime scenes to mtDNA from given individuals to ascertain whether the tested individuals are within the maternal line (or another coincidentally matching maternal line) of people who could have been the source of the trace evidence.

Analytical Methods and Categorical Matches

The small mitochondrial genome can be analyzed with sequencing methods or with microarrays that give the order of some or all the base pairs.²⁵⁸ The traditional Sanger-sequencing technology usually was done on just the hypervariable parts of the coding region. Cheaper and faster massively parallel sequencing is replacing that procedure, increasing the ability of laboratories to detect heteroplasmy and better distinguish among different maternal lineages in the population.²⁵⁹ The laboratory first generates descriptions of the sequences of two

253. See, e.g., *Silent Witness: Forensic DNA Analysis in Criminal Investigations and Humanitarian Disasters* (Henry Ehrlich et al. eds., 2020).

254. “Length heteroplasmy” (most often in the form of simple repeats such as CC . . .) is common but harder to characterize precisely in sequencing studies.

255. Alison Barrett et al., *Pronounced Somatic Bottleneck in Mitochondrial DNA of Human Hair*, 375 *Phil. Trans. Roy. Soc’y B* 20190175, at 2 (2019), <https://doi.org/10.1098/rstb.2019.0175>.

256. See *infra* section titled “Population Genetics and Statistics.”

257. Estimates of the average mutation rate range up to around 1 per 70 generations. Mikkel M. Andersen & David J. Balding, *How Many Individuals Share a Mitochondrial Genome?*, 14 *PLoS Genetics* e1007774 (2018), <https://doi.org/10.1371/journal.pgen.1007774>.

258. See *supra* sections titled “Sequence-Specific Probes and Microarrays” and “Sequencing and SNPs.”

259. For a survey of the major methods, see Bethany Forsythe et al., *Methods for the Analysis of Mitochondrial DNA*, 3 *WIREs Forensic Sci.* e1388 (2021), <https://doi.org/10.1002/wfs2.1388>.

samples—say, DNA extracted from a hair shaft found at a crime scene and hairs plucked from a suspect.²⁶⁰ Let’s assume that there is no issue as to whether the technology has been properly applied with suitable controls,²⁶¹ and the factfinder can be confident that the sequences as reported truly represent the order of the base pairs in the mtDNA of the hairs. What should the factfinder make of the two sequences?

Most analysts use match/no-match categories to describe the result of a comparison.²⁶² The Department of Justice currently favors the words “[c]annot be excluded (i.e., inclusion, or included) [and] [e]xclusion (i.e., excluded).”²⁶³ As will be explained in the section below titled “Heteroplasmy,” an intermediate category of “inconclusive” covers less-than-perfectly-matching sequences.²⁶⁴ In the simplest cases, the two sequences show a good number of differences (a clear exclusion), or they are identical (an inclusion). Suppose there is an exclusion. After explaining why a clear “exclusion” means that such a large difference almost certainly cannot occur for two hairs from people within the same lineage, an analyst’s testimony on direct examination could end. But if the analyst were to stop at the same point when declaring an inclusion, the factfinder would be left with no scientific information on how common or rare such matching sequences are for different lineages in the relevant population. Such inclusion-only testimony invites the objection that the congruent sequences are of unknown probative value and might be given more weight than they deserve.²⁶⁵ To more

260. Mitotyping often can exclude individuals as the source of stray hairs even when the hairs are microscopically indistinguishable. Nonetheless, gross or microscopic hair comparison can be useful, as a screening test, to exclude a suspect without the need for DNA analysis. See ASTM E3316–22, Standard Guide for Forensic Examination of Hair by Microscopy (2022).

261. See Walther Parson et al., *DNA Comm’n of the Int’l Soc’y for Forensic Genetics: Revised and Extended Guidelines for Mitochondrial DNA Typing*, 13 *Forensic Sci. Int’l: Genetics* 134, 135–36 (2014), <https://doi.org/10.1016/j.fsigen.2014.07.010>; Scientific Working Group on DNA Analysis Methods, SWGDAM Interpretation Guidelines for Mitochondrial DNA Analysis by Forensic DNA Testing Laboratories § 1 (Apr. 23, 2019) [hereinafter SWGDAM MtDNA Guidelines].

262. E.g., *Lewis v. State*, 889 So.2d 623, 673 (Ala. Ct. Crim. App. 2003) (FBI expert testified that “[t]he fourth step is to compare the order of the bases, i.e., the sequence, to other samples to determine if there is a ‘match.’”); see also SWGDAM MtDNA Guidelines, *supra* note 261, § 3.1.

263. U.S. Dep’t of Just., Uniform Language for Testimony and Reports for Forensic Mitochondrial DNA Examinations 2 (Dec. 12, 2022), <https://perma.cc/J6V4-Z4EG> [hereinafter MtDNA ULTR]. This terminology document does not single out “match” or words based on that stem as impermissible. It defines these labels as “an examiner’s conclusion” with no specification for how the examiner should reach these conclusions.

264. In principle, a likelihood ratio is better suited to cases in which there are slight sequence differences (and could be used in all situations). See Coble et al., *supra* note 233, at 319, 324–41; cf. *supra* section titled “Likelihood ratios (LRs)” (on likelihood ratios for Y-STRs).

265. But see Commonwealth v. Chmiel, 30 A.3d 1111, 1158 (Pa. 2011) (suggesting that “statistical analysis is . . . unnecessary”).

adequately inform the court or jury of the implications of a match, forensic scientists invoke principles of population genetics and statistics.²⁶⁶

Population Genetics and Statistics for Mitochondrial DNA

Population databases

As with Y-STRs, to indicate the significance of the match, analysts usually estimate the frequency of the sequence in some population. It is not necessary to combine any allele frequencies because the entire mtDNA sequence, whatever its internal structure may be, is inherited as a single unit (a haplotype). In other words, the sequence itself is like a single allele, and one can simply count how often it occurs in a sample of unrelated people.²⁶⁷

Laboratories therefore refer to databases of mtDNA sequences to see how often the type in question has been seen before and to compute the probability of a matching haplotype in a random draw from the population given that it has been observed in the case under investigation.²⁶⁸ The issues are essentially the same as those for Y haplotypes.²⁶⁹ Key questions are which reference population or populations to use;²⁷⁰ what estimator to use for the sequence frequency or match probability;²⁷¹ how to account for sampling error in this estimate;²⁷²

266. The Department of Justice expects analysts in its laboratories to “provide a quantitative statement describing the weight of the evidence for all inclusions regardless of the magnitude of the resulting quantitative value.” MtDNA ULTR, *supra* note 263, at 3.

267. In this context, “unrelated people” means individuals with a different maternal lineage.

268. The publicly available European DNA Profiling Group (EDNAP) mtDNA Population Database (EMPOP), which houses quality-controlled sequence data from populations across the world, was established in 2006. Walther Parson & Arne Dür, *EMPOP—A Forensic mtDNA Database*, 1 *Forensic Sci. Int'l: Genetics* 88 (2007), <https://doi.org/10.1016/j.fsigen.2007.01.018>.

269. See *supra* section titled “Population Genetics and Statistics.”

270. See Parson et al., *supra* note 261, at 140 (“Local, regional and continental databases may all be searched and reported to guide evaluation of the evidence. Additionally, frequencies from geographic databases of relevant differentiated populations may be reported (as in the United States where reports often list separately search results from major population groups).”); *supra* section titled “Haplotype frequency or probability estimates.”

271. See Parson et al., *supra* note 261, at 140; *supra* section titled “Haplotype frequency or probability estimates” (noting most of the alternatives). In *State v. Pappas*, 776 A.2d 1091 (Conn. 2001), the defendant maintained that the usual sample proportion is inadmissible because an estimate that tosses in the haplotype in the case itself is more appropriate. The Connecticut Supreme Court responded that “using zero as the numerator rather than one . . . goes to the weight of the evidence and not to its admissibility.” *Id.* at 1109.

272. See Parson et al., *supra* note 261, at 140 (“approaches that empirically take into account the haplotype distribution of the database in question (e.g. the Kappa model . . .) may best represent the strength of the mtDNA evidence”); SWGDAM MtDNA Guidelines, *supra* note 261, § 4.2.4 (binomial distribution); *supra* section titled “Haplotype frequency or probability estimates”; *infra* note 275.

whether and how to adjust for population structure;²⁷³ whether to present likelihood ratios, including one that combines the information in mitochondrial haplotypes with that from Y haplotypes (or both Y and autosomal haplotypes);²⁷⁴ and how to present or explain the statistical assessment to a factfinder.²⁷⁵

Heteroplasmy

The simple inclusion–exclusion approach must be modified to account for the fact that the same individual can have detectably different (albeit very similar) mitotypes in different tissues or even in different cells in the same tissue. To understand the implications of heteroplasmy, we need to understand how it comes into existence. Heteroplasmy can occur because of mtDNA mutations during the division of adult cells, such as those at the roots of hair shafts. These new mitotypes are confined to the individual. They will not be passed on to future generations. Heteroplasmy also can result from a mutation contained in the egg cell that grew into an individual. Such mutations can make their way into succeeding generations, establishing new mitotypes in the population. But this is an uncertain process. Egg cells contain many mitochondria, and the

273. See *supra* section titled “Haplotype frequency or probability estimates.” Compare Coble et al., *supra* note 233, at 334–35 (describing a procedure) with SWGDAM MtDNA Guidelines, *supra* note 261, § 4.3 (“However, determination of an appropriate theta (θ) value is complicated by the variety of primer sets, covering different portions of HV1 and/or HV2, which may be applied to forensic casework. SWGDAM has not yet reached consensus on the appropriate statistical approach to estimating θ for mtDNA comparisons.”).

274. See Parson et al., *supra* note 261, at 140; SWGDAM MtDNA Guidelines, *supra* note 261, §§ 4.4 & 4.5.

275. Apparently, SWGDAM would dispense with estimates of sampling error. SWGDAM MtDNA Guidelines, *supra* note 261, § 4.2.3 (“Reporting an mtDNA haplotype sample frequency without a confidence interval is acceptable as a factual statement regarding observations in the database.”). But the conclusion that judges or jurors can make suitable use of a sample proportion without an assessment of statistical uncertainty is not a strictly scientific judgment. A court might need to consider whether it is sufficient to leave it to the jury to decide how to weigh the fact of the match and the absence of the same sequence in a convenience sample that might—or might not—be representative of the local population.

As with Y haplotyping (*supra* note 227), the use of exclusion probabilities that do not account for inconclusives could be misleading. In *Vaughn v. State*, 646 S.E.2d 212, 215 (Ga. 2007), for instance, the statement that a suspect “cannot be excluded” when “there is a single base pair difference” was transformed into “a match.” These inconclusive sequences contribute to the number of people who would not be excluded. Therefore, in *Pappas*, it was misleading to conclude “that approximately 99.75% of the Caucasian population could be excluded as the source of the mtDNA in the sample.” 776A.2d 1091, 1104 (Conn. 2001) (footnote omitted). This percentage neglects the individuals whose mtDNA sequences are off by one base pair. Along with the 0.25% who are included because their mtDNA matches completely, these one-off people would not be excluded. Of course, the difference may be fairly small. In *Pappas*, a defense expert reported that the actual nonexclusion rate was still “99.3 percent of the Caucasian population.” *Id.* at 1105 (footnote omitted).

mature egg cell will not contain just the mutation—it will house a mixed population of the old-style mitochondria and a number of the mutated ones (with DNA that usually differs from the original at a single base pair). Figuratively speaking, the original mtDNA sequence and the mutated version fight it out for several generations until one of them becomes “fixed” in the population. In the interim, the progeny of the mutated egg cell will harbor both strains of mitochondria.

When mtDNA from a single-source crime-scene sample is compared to a suspect’s sample, there are three possibilities: (1) neither sample is detectably heteroplasmic; (2) one sample displays heteroplasmy, but the other does not; (3) both samples display heteroplasmy. In each scenario, the comparison can produce an exclusion or an inclusion:

1. *Neither sample heteroplasmic.* In the first situation, if the sequence in the crime-scene sample is markedly different from the sequence in the suspect’s sample, then the suspect is excluded. But heteroplasmy could be the reason for a difference of only a single base or so. For example, the sequence in a hair shaft coming from the suspect could be a slight mutation of the dominant sequence in the suspect. Therefore, the FBI treats a difference at a single base pair as inconclusive.²⁷⁶ When the one mtDNA sequence characteristic of each sample is identical, the issue becomes how to use the reference database of mtDNA sequences, as discussed above.
2. *Suspect’s sample heteroplasmic, crime-scene sample not.* One version of the second scenario arises when heteroplasmy is seen in the suspect’s tissues but not in the crime-scene sample. If the crime-scene sequence is not close to either of the suspect’s sequences, then the suspect is excluded. If it is identical to one of the suspect’s sequences, then the suspect is included, and a suitable reference database should indicate how infrequent such an inclusion would be. If crime-scene DNA is one base pair removed from either of the suspect’s sequences, then the result is inconclusive.
3. *Both samples heteroplasmic.* In this third scenario, multiple sequences are seen in each sample. To keep track of things, we can call the sequences in the crime-scene sample C1 and C2, and those in the suspect’s sample S1 and S2. If either C1 or C2 is very different from both S1 and S2,

276. SWGDAM MtDNA Guidelines, *supra* note 261, § 3.1. Where did the one-base-pair rule come from? With traditional Sanger sequencing of parts of the control region of the mitogenome, heteroplasmy at a single point seems to occur, on average, in approximately 6% of the control-region sequences from around the world; heteroplasmy at multiple sites is seen in less than 1% of sequences. Jodi A. Irwin et al., *Investigation of Heteroplasmy in the Human Mitochondrial DNA Control Region: A Synthesis of Observations from More Than 5000 Global Population Samples*, 68 J. Molecular Evolution 516 (2009), <https://doi.org/10.1007/s00239-009-9227-4>.

the suspect is excluded. If C1 and C2 are the same as S1 and S2, the suspect is included. Because detectable heteroplasmy is not very common, this inclusion is stronger evidence of identity than the simple match in the first scenario. Finally, in some of the situations in which C1 or C2 are merely very close to S1 or S2, the comparison will be inconclusive.

Emerging Questions for Courts

A number of courts have rejected objections that Sanger sequencing of hyper-variable regions of the mitochondrial DNA genome does not comport with *Frye*²⁷⁷ or *Daubert*.²⁷⁸ But forensic-DNA laboratories are adopting some of the massively parallel sequencing (MPS) methods, described in the section above titled “Sequencing methods,” that can work with smaller quantities of mtDNA and that can efficiently sequence the entire mitogenome.²⁷⁹ The widespread use of MPS methods in clinical medicine and genetic research indicates scientific acceptance and validity at a general level (see “Sequencing methods”), but courts still may encounter questions about the rate of errors in the “reads,” what steps are taken to correct them, precautions against misinterpreting nuclear mtDNA segments (NUMTs) as present in mitochondria, and more. Much depends on the equipment and its operation.²⁸⁰ Published validity studies from biotechnology companies developing the sequencers and kits, from the individual forensic

277. *E.g.*, *Magaletti v. State*, 847 So. 2d 523, 528 (Fla. Dist. Ct. App. 2003) (“[T]he mtDNA analysis conducted [on hair] determined an exclusionary rate of 99.93 percent. In other words, the results indicate that 99.93 percent of people randomly selected would not match the unknown hair sample found in the victim’s bindings.”); *People v. Sutherland*, 860 N.E.2d 178, 271–72 (Ill. 2006); *People v. Holtzer*, 660 N.W.2d 405, 411 (Mich. Ct. App. 2003); *Wagner v. State*, 864 A.2d 1037, 1043–49 (Md. Ct. Spec. App. 2005) (mtDNA sequencing admissible despite contamination and heteroplasmy).

278. *E.g.*, *United States v. Beverly*, 369 F.3d 516, 531 (6th Cir. 2004) (“The scientific basis for the use of such DNA is well established.”); *United States v. Coleman*, 202 F. Supp. 2d 962, 967 (E.D. Mo. 2002) (“[a]t the most, seven out of 10,000 people would be expected to have that exact sequence of As, Ts, Cs, and Gs”), *aff’d*, 349 F.3d 1077 (8th Cir. 2003); *Pappas*, 776 A.2d at 1095; *State v. Underwood*, 518 S.E.2d 231, 240 (N.C. Ct. App. 1999); *State v. Council*, 515 S.E.2d 508, 518 (S.C. 1999); *State v. Griffin*, 384 P.3d 186, 202 (Utah 2016).

279. The methods being implemented are not the most advanced ones that have been proposed for mitochondrial research. It is technically possible to zoom in on the mitogenome (as a single molecule, without PCR amplification) and sequence it in a single “read.” Ieva Keraite et al., *A Method for Multiplexed Full-Length Single-Molecule Sequencing of the Human Mitochondrial Genome*, 13 *Nature Communications* 5902 (2022), <https://doi.org/10.1038/s41467-022-33530-3> (using CRISPR-Cas9 and claiming to overcome the relatively large rate of read errors in single-molecule sequencing systems).

280. See, *e.g.*, Yun Heo et al., *Comprehensive Evaluation of Error-Correction Methodologies for Genome Sequencing Data*, in *Bioinformatics* (Nakaya Helder ed., 2021); Nicholas Stoler & Anton

laboratories deploying them, and from academic researchers can be scrutinized.²⁸¹ At least in early admissibility hearings, courts may wish to consider appointing independent expert advisers for this purpose.

Questions of the accuracy of the sequences and their interpretation also might emerge with respect to the “reliably applied” prong of the requirements for scientific evidence under Rule 702. The presence of and adherence to rigorous quality control and assurance systems might be significant.²⁸² Moreover, all these matters are potential topics that can arise for jurors (and judges when they act as triers of fact). The practice followed in a case at bar can be compared to both minimally acceptable and recommended practices as articulated in the scientific literature and guidelines from expert groups.²⁸³

Courts have consistently held that neither the phenomenon of heteroplasmy nor the limitations in the statistical analysis preclude the forensic use of sequencing results under either Rule 702 or Rule 403.²⁸⁴ However, mitogenome MPS picks up more heteroplasmic sites than Sanger sequencing, and more individuals exhibit not just one, but several points of heteroplasmy.²⁸⁵ As a result, restricting the “inconclusive” category to only a single sequence–difference across the mitogenome will not prevent as many false exclusions. Despite the allure of purportedly definitive “exclusions” and “inclusions,” the gray zone of “inconclusive” sequence differences is increasingly hard to define with a simple rule of thumb for the entire mitogenome. Courts may need to inquire into what the literature demonstrates about the accuracy of the laboratory’s rule or procedure for assessing the extent to which sequence similarities and differences support conclusions (including judgments of “inconclusive”) about the maternal relatedness of the unknown individual whose mtDNA was recovered and the suspect’s comparison sample.

Nekrutenko, *Sequencing Error Profiles of Illumina Sequencing Instruments*, 3 *Nucleic Acids Resch. Genomics & Bioinformatics* lqab019 (2021), <https://doi.org/10.1093/nargab/lqab019>.

281. E.g., Michael D. Brandhagen et al., *Validation of NGS for Mitochondrial DNA Casework at the FBI Laboratory*, 44 *Forensic Sci. Int’l: Genetics* 102151 (2020), <https://doi.org/10.1016/j.fsigen.2019.102151> (sequencing the control region); Jennifer Churchill Cihlar et al., *Developmental Validation of a MPS Workflow with a PCR-Based Short Amplicon Whole Mitochondrial Genome Panel*, 11 *Genes* 1345 (2020), <https://doi.org/10.3390/genes11111345> (whole-mitogenome sequencing).

282. See *supra* section titled “Sample Collection and Laboratory Performance.”

283. See *supra* note 261.

284. E.g., *Beverly*, 369 F.3d at 531 (“[T]he mathematical basis for the evidentiary power of the mtDNA evidence was carefully explained, and was not more prejudicial than probative.”); *Pappas*, 776 A.2d 1091; *Griffin*, 384 P.3d at 202–03.

285. Jennifer A. McElhoe et al., *Exploring Statistical Weight Estimates for Mitochondrial DNA Matches Involving Heteroplasmy*, 136 *Int’l J. Legal Med.* 671, 695–96 (2022), <https://doi.org/10.1007/s00414-022-02774-5>.

Mixtures

What Are DNA Mixtures?

Samples of biological trace evidence recovered from crime scenes often contain a mixture of fluids or tissues from different individuals. Examples include vaginal swabs collected as sexual assault evidence, bloodstain evidence from scenes where several individuals shed blood, and objects that several people touched. However, not all mixed samples produce mixed STR profiles. Consider a sample in which 99% of the DNA comes from the defendant and 1% comes from a different individual. Even if some of the molecules from the minor contributor come in contact with a primer and the polymerase and an STR is amplified, the resulting signal might be too small to detect—the peak in an electropherogram will blend into the background. Because the vast bulk of the amplified STRs will come from the defendant's DNA, the electropherogram may show only one STR profile. In these situations, the interpretation of the single DNA profile is the same as when 100% of the DNA molecules in the sample are the defendant's.

When the mixtures are more evenly balanced among contributors, however, the STRs from multiple contributors can appear as “extra” peaks. As a rule, because DNA from a single individual can have no more than two alleles at each locus,²⁸⁶ the presence of three or more peaks at several loci indicates a mixture of DNA in the sample.²⁸⁷ Figure 6 shows another electropherogram from DNA recovered in *People v. Pizarro*.²⁸⁸ The fact that there are as many as four alleles at

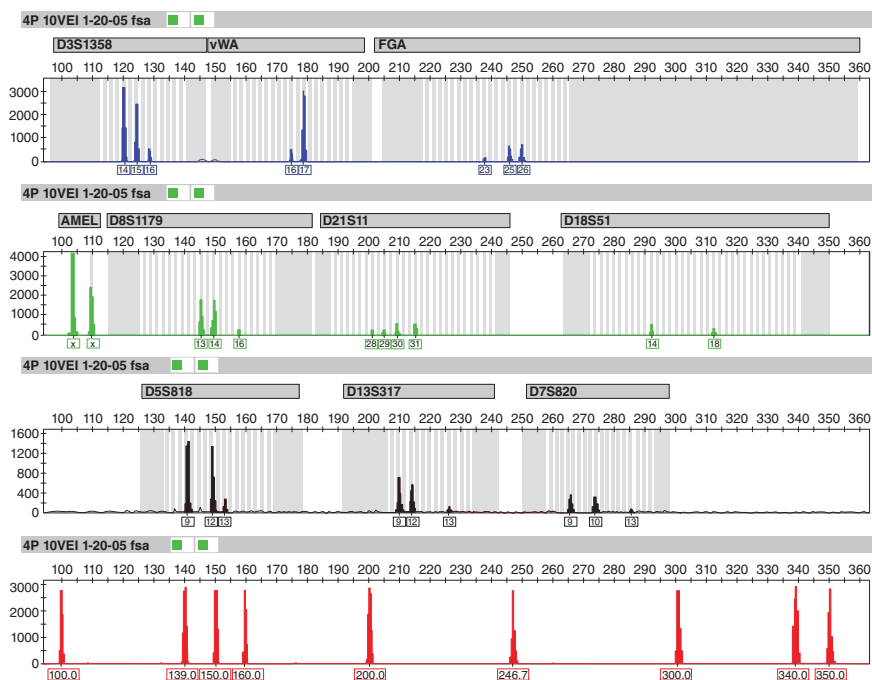
286. This follows from the fact that individuals inherit chromosomes in pairs, one from each parent. See *supra* section titled “Sexual Reproduction and the Genome.” An individual who inherits the same allele from each parent (a homozygote) can contribute only that one allele to a sample, and an individual who inherits a different allele from each parent (a heterozygote) will contribute those two alleles. Finding three or more alleles at several loci therefore indicates a mixture.

287. On rare occasions, an individual exhibits a phenotype with three alleles at a locus. This can be the result of a chromosome anomaly (such as a duplicated gene on one chromosome or a mutation). A sample from such an individual is usually easily distinguished from a mixed sample. The three-allele variant is seen at only the affected locus, whereas with mixtures, more than two alleles typically are evident at several loci.

Individuals who are genetic chimeras also can produce DNA samples that could be mistaken for that of two different people. Chimerism refers to more than one cell line resulting from different zygotes (fertilized eggs). Chimerism can arise artificially, as a result of a blood transfusion or transplantation, or spontaneously, during the formation and development of an embryo. Chimerism: A Clinical Guide 3–48 (Nicole L. Draper ed., 2018). The implications for forensic-DNA identification are discussed in, e.g., David H. Kaye, *Chimeric Criminals*, 14 Minn. J.L. Sci. & Tech. 1 (2012); Erin Murphy, *Inside the Cell: The Dark Side of Forensic DNA* 246–49 (2015); Kayla Sheets & Robert Wenk, *Relationship Testing and Forensics*, in Chimerism: A Clinical Guide, *supra*, at 51–63.

288. See *supra* Figure 4.

Figure 6. Electropherogram in *People v. Pizarro* that can be interpreted as a mixture of DNA from the victim and the defendant.



Source: Steven Myers and Jeanette Wallin, California Department of Justice, provided the image.

That some loci in Figure 6, such as vWA, exhibit only two peaks does not preclude the existence of a two-person mixture. Both the victim and the rapist (on the state's theory of the case) might share the same alleles at that locus, and the amplified product of the victim's DNA might be masking that from the rapist. Both might be homozygous at that locus (each producing one peak). If it is credible that one or both alleles of the rapist's vWA locus dropped out of the electropherogram (not likely with substantial quantities of DNA), still more scenarios consistent with the state's theory of the case could explain why there are only two peaks at the vWA locus. Enumerating and assessing such possibilities leads to the topic of "mixture deconvolution."

some loci and that many of the peaks match the victim's suggests that the sample is a mixture of the victim's and another person's DNA. Furthermore, the presence of two peaks at the amelogenin locus (labeled AMEL in the electropherogram) shows that male DNA is part of the mixture. Because all the peaks that do not match the victim are part of the defendant's STR profile, the mixture is consistent with the state's theory that the defendant raped the victim.

Have Strategies for Simplifying Mixtures Interpretation Been Pursued?

A pioneer in the development of DNA typing methods and their interpretation had this advice for forensic laboratories: “Don’t do mixture interpretations unless you have to.”²⁸⁹ If a laboratory has other samples that do not show evidence of mixing, it may be able to avoid fully deciphering mixed profiles. Even across a single stain, the proportions of a mixture can vary, and it might be possible to extract a DNA sample that does not produce a mixed profile.

In sexual assault cases, it sometimes is possible to separate sperm from the male and female epithelial cells on vaginal, oral, or rectal swabs and then analyze the sperm DNA by itself.²⁹⁰ When this procedure works well, and only one man’s sperm are present, the electropherogram should display one clear profile with little or no interference from the female DNA. And even with sperm from more than one man, the analysis might be simpler because the victim’s DNA no longer can mask alleles from the sperm. Also, in sexual assault (and other) cases, Y chromosome testing can reveal the number of men (from different paternal lines) whose DNA is being detected and whether the defendant’s Y chromosome is consistent with one of these paternal lines.²⁹¹ Because only males have Y chromosomes, the female DNA in a mixture has no effect.

Finally, rather than extracting DNA from a mixture of cells and performing PCR–CE on the resulting mixture of DNA molecules, it may be possible to isolate individual cells and analyze the DNA one cell at a time,²⁹² either by pushing

289. Peter Gill, as quoted in John M. Butler, *Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers* 166 (2d ed. 2005).

290. The nucleus of a sperm cell lies behind a protective structure that does not break down as easily as the membrane in an epithelial cell. This makes it possible to disrupt the male and female epithelial cells first and remove their DNA, and then to use a harsher treatment to disrupt the sperm cells. Heather E. McKiernan & Phillip B. Danielson, *Molecular Diagnostic Applications in Forensic Science*, in *Molecular Diagnostics* 371 (George P. Patrinos ed., 3d ed. 2017). Other methods of differential extraction also have been studied or applied to forensic samples. E.g., Lana Ostojic et al., *Micromanipulation of Single Cells and Fingerprints for Forensic Identification*, 51 *Forensic Sci. Int’l: Genetics* 102430 (2021), <https://doi.org/10.1016/j.fsigen.2020.102430>.

291. E.g., *State v. Polizzi*, 924 So. 2d 303, 308–09 (La. Ct. App. 2006) (testing for Y-STRs on “the genital swab with the DNA profile from the Defendant’s buccal swab, . . . the Defendant or any of his paternal relatives could not be excluded as having been a donor to the sample from the victim,” while “99.7 percent of the Caucasian population, 99.8 percent of the African American population, and 99.3 percent of the Hispanic population could be excluded as donors of the DNA in the sample”); *State v. Bander*, 208 P.3d 1242, 1250 (Wash. Ct. App. 2009) (Y-STR profiling showed that “99.9 percent of the population could be excluded as possible contributors to the sample extracted from the electrical cords recovered from Gardner’s body”).

292. See Jianye Ge et al., *Precision DNA Mixture Interpretation with Single-Cell Profiling*, 12 *Genes* 1649 (2021), <https://doi.org/10.3390/genes12111649>.

PCR–CE to its limits²⁹³ or by employing the newer sequencing methods sketched in the section above titled “Sequencing methods.”²⁹⁴ Without physical separation, however, the multiple genotypes in the sample can produce mixed profiles. Analysts then have to interpret these more complex patterns in terms of the individual genotypes that might be present. Inferring the probable genotypes that are combined in a mixed profile is known as mixture deconvolution. This interpretive process is described in the next subsection.

What Is Mixture Deconvolution?

The goal of mixture deconvolution is twofold: (1) to deduce the individual genotypes that are intertwined in an electropherogram that contains peaks from the DNA of different individuals; and (2) to assess the significance of the fact that a defendant has some or all of such a genotype.²⁹⁵ This analysis is computational. A simplified example illustrates the nature of the procedure that has been in use for over 20 years.²⁹⁶ Suppose we create a mixture by adding equal, easily detectable amounts of DNA from two unrelated individuals and type it at just one autosomal STR locus. Let us say that the contributors are each heterozygotes—each individual has two alleles per locus—and that all the alleles are distinct. The electropherogram will have four peaks of roughly equal heights at locations that we can designate A, B, C, and D. An analyst who is given this electropherogram without any information about the contributors could reason as follows:

293. E.g., Kaitlin Huffman et al., *Y-STR Mixture Deconvolution by Single Cell Analysis*, 68 J. Forensic Sci. 275 (2023), <https://doi.org/10.1111/1556-4029.15150> (individual Y-STR profiles recovered from direct PCR amplification of single-cell subsamples from two- and three-person mixtures).

294. See also section titled “Did the Sample Contain Enough DNA?” and note 112 above; cf. Lucie Kulhankova et al., *Single-Cell Transcriptome Sequencing Allows Genetic Separation, Characterization and Identification of Individuals in Multi-Person Biological Mixtures*, 6 Communications Biology 201 (2023), <https://doi.org/10.1038/s42003-023-04557-z> (initial study using single-cell RNA transcript sequences and other genetic data to successfully deconvolute laboratory-created mixtures of blood from 2- to 9-person mixtures in ratios as low as 1 to 60).

295. Some courts have held mixture interpretations admissible without any statistical assessment. E.g., *Rodriguez v. State*, 273 P.3d 845, 851–52 (Nev. 2012); *People v. Her*, 157 Cal. Rptr. 3d 40 (Cal. Ct. App. 2013).

296. Tim M. Clayton et al., *Analysis and Interpretation of Mixed Forensic Stains Using DNA STR Profiling*, 91 Forensic Sci. Int'l 55 (1998), [https://doi.org/10.1016/s0379-0738\(97\)00175-1](https://doi.org/10.1016/s0379-0738(97)00175-1); Ian W. Evett et al., *Taking Account of Peak Areas When Interpreting Mixed DNA Profiles*, 43 J. Forensic Sci. 62 (1998); Peter Gill et al., *Interpreting Simple STR Mixtures Using Allele Peak Areas*, 91 Forensic Sci. Int'l 41 (1998), [https://doi.org/10.1016/s0379-0738\(97\)00174-6](https://doi.org/10.1016/s0379-0738(97)00174-6).

The minimum number of contributors is two, but there could be more. The four peaks are of equal height, so if there are only two contributors to the mixture, their genotypes could be any of the following six combinations:

CONTRIBUTOR 1	CONTRIBUTOR 2
AB	CD
AC	BD
AD	BC
CD	AB
BD	AC
BC	AD

These can be written more compactly using a slash to separate the two contributors' genotypes and listing them as AB/CD, AC/BD, etc. There cannot be exactly three contributors (unless one of them did not contribute enough DNA to have any impact on the peak heights).²⁹⁷ If there are four contributors, the genotypes could be AB/CD/AB/CD, AC/BD/AC/BD, AD/BC/AD/BC, AA/BB/CC/DD, AC/BC/AB/DD, AC/BC/AD/BD, AC/AD/CD/BB, or BC/BD/CD/AA and more (rearranging the order to assign the two-allele genotypes to the different contributors 1 through 4). As five or more contributors are considered, it gets still more complicated.

If the mixed, two-contributor sample had markedly different proportions of DNA, there could be a pattern of tall and short peaks across multiple loci (as in Figure 6), suggesting that the short ones represent the DNA of a “minor contributor” and the tall ones come from the DNA of a “major contributor.” By using the peak heights and locations, the analyst might be able to “deconvolute” the pattern of peaks into the most likely underlying genotypes and then see if the defendant has one of these genotypes.²⁹⁸ Great effort has gone into formulating a systematic procedure for determining the number of contributors to consider,²⁹⁹ to listing all possible sets of genotypes that could give rise to a mixed STR profile, to scratching some off the list as inconsistent with the peak-height information, and to performing a statistical assessment of an inclusion of a defendant who has one of the genotypes left on the list. The standard algorithm for this “manual” deconvolution has six or seven steps.³⁰⁰ The process requires the

297. For instance, if the three genotypes were AB/CD/AB, the A and B peaks would be twice the height of the C and D peaks, but they are not.

298. See, e.g., *Roberts v. United States*, 916 A.2d 922, 932–35 (D.C. 2007) (holding such inferences to be admissible).

299. Decisions or assumptions about the number of contributors could be assisted by other evidence about the events that resulted in the mixture.

300. For summaries of the steps, see Jo-Anne Bright & Michael D. Coble, *Forensic DNA Profiling: A Practical Guide to Assigning Likelihood Ratios* 7–15 (2020); Butler et al., *supra* note 105; see also Frederick R. Bieber et al., *Evaluation of Forensic DNA Mixture Evidence: Protocol for Evaluation, Interpretation, and Statistical Calculations Using the Combined Probability of Inclusion*, 17 BMC Genetics 125

analyst to make mostly binary decisions about which peaks are alleles as opposed to artifacts (stutter peaks, pull-up, dye blobs).³⁰¹

It is now generally recognized that to avoid an “expectation effect” on the analyst’s judgments, the allele calls should be made before considering the genotypes of any suspects or other possible contributors to the mixed sample. Then the analyst must decide which genotypes are included or excluded from the mixture. The task becomes even more difficult when the quantity of the DNA from some contributor is marginal, with the possibility of some peaks that would be expected for larger amounts of DNA dropping out to leave only a partial profile.³⁰² Mixtures of degraded or inhibited DNA also are harder to interpret because parts of individual profiles may be missing. Nevertheless, it has been said that “[t]he binary method served the forensic community well for a number of years; however it was mostly restricted to single-source, two-person, and some three-person mixtures (e.g., those with a clear major component).”³⁰³ Other observers saw manual mixture interpretation as erratic,³⁰⁴ with some laboratories

(2016), <https://doi.org/10.1186/s12863-016-0429-7>; Peter Gill et al., *DNA Commission of the International Society of Forensic Genetics: Recommendations on the Interpretation of Mixtures*, 160 *Forensic Sci. Int’l* 90 (2006), <https://doi.org/10.1016/j.forsclint.2006.04.009>; Scientific Working Group on DNA Analysis Methods (SWGDM), *SWGDM Interpretation Guidelines for Autosomal STR Typing by Forensic DNA Testing Laboratories* (Jan. 12, 2017, rev. July 13, 2021), <https://perma.cc/2ZCW-Q6L6>.

301. On these artifacts, see *supra* section titled “Capillary electrophoresis and fluorescence.”

302. On drop-out, see *supra* section titled “Did the Sample Contain Enough DNA?” For a detailed analysis of the manual interpretation steps in the context of an example with these complications, see Michael D. Coble, *Worked Mixture Example*, in Butler, *supra* note 113, at 537.

303. Bright & Coble, *supra* note 300, at 15.

304. An oft-cited but limited study is Itiel E. Dror & Greg Hampikian, *Subjectivity and Bias in Forensic DNA Mixture Interpretation*, 51 *Sci. & Just.* 204 (2011), <https://doi.org/10.1016/j.scijus.2011.08.004>. It reports aggregated conclusions of (1) two analysts in Georgia who compared a complex four- or five-man mixture in a gang-rape case with the profiles of three suspects, and (2) 17 analysts at an unnamed laboratory given, as an abstract exercise, the mixed profile and the profile of just one of the three suspects. The two groups reached different conclusions. The extent to which the reported differences prove much about the impact of biasing information and the reproducibility of manual mixture deconvolution in general is discussed in David H. Kaye, *The Design of “The First Experimental Study Exploring DNA Interpretation,”* 52 *Sci. & Just.* 126 (2012), <https://doi.org/10.1016/j.scijus.2011.10.003>, and Itiel E. Dror, *Cognitive Forensics and Experimental Research About Bias in Forensic Casework*, 52 *Sci. & Just.* 128 (2012), <https://doi.org/10.1016/j.scijus.2012.03.006>. More systematic interlaboratory comparisons of up to five-person mixtures from 1997 through 2018 are tabulated in John M. Butler et al., *DNA Mixture Interpretation: A NIST Scientific Foundation Review* 80–81 (NISTIR 8351-DRAFT, June 2021) (tbl. 4.8), <https://doi.org/10.6028/NIST.IR.8351-draft>. The reasons for variability in the reports from participants in two U.S. interlaboratory studies are outlined in John M. Butler et al., *NIST Interlaboratory Studies Involving DNA Mixtures (MIX05 and MIX13): Variation Observed and Lessons Learned*, 37 *Forensic Sci. Int’l: Genetics* 80, 81 (2018), <https://doi.org/10.1016/j.fsigen.2018.07.024>; cf. Michael Coble et al., *Analysis of Forensic Mixtures*, in *Forensic DNA Analysis in Criminal Investigations and Humanitarian Disasters* 49, 58–59 (Henry Ehrlich et al. eds., 2020) (“Studies from European laboratories [citations omitted] have typically shown wide variation in the results within and between

performing the task poorly.³⁰⁵ Despite the degree to which “subjective” judgment can go into the determination of which peaks are real, which are artifacts, which are masked, and which are absent for some other reason, courts generally have rejected arguments that manual mixture analysis is so unreliable or so open to manipulation that the results are inadmissible.³⁰⁶

What Statistical Assessment of the Comparison to a Defendant’s Profile Has Been Conducted?

When manual deconvolution of a mixed profile is complete and a defendant’s profile is compatible with it, most laboratories will attempt to supply a statistical assessment to indicate how powerful the evidence is.³⁰⁷ Methods for doing so fall

the participants’ laboratories. [In] two interdisciplinary studies in 2005 and 2013 . . . involving laboratories performing STR genotyping in the United States and Canada . . . overall results were concordant with the studies from Europe. . . .”). Proficiency tests, which sometimes include two- or three-person mixtures, are listed in Butler et al., *DNA Mixture Interpretation: A Scientific Review*, *supra*, at 78–79 (tbl. 4.7) (reporting small error rates); *see also* Todd Bille et al., *Study of CTS DNA Proficiency Tests with Regard to DNA Mixture Interpretation: A NIST Scientific Foundation Review*, 13 *Genes* 2171 (2022), <https://doi.org/10.3390/genes13112171> (re-analyzing the proficiency test data and concluding that “[s]ample switching, contamination, and reporting results incorrectly are serious errors [but not a consequence of] the mixture interpretation strategy”).

305. For example, an inquiry into mixture interpretations at the Washington, D.C., forensic science laboratory at the behest of the U.S. Attorney’s Office for the District of Columbia led to the ouster of the laboratory’s management. *See supra* note 129. The common errors that were identified mostly pertained to the statistical interpretation of inclusions in mixture cases. Barber v. United States, 179 A.3d 883, 892 n.17 (D.C. 2018). Other jurisdictions have issued letters to convicted defendants notifying them that older calculations of the combined probability of inclusion (*see infra* section titled “(1) Combined Probability of Inclusion (CPI) and Exclusion (CPE)”) might be inaccurate. Skinner v. State, 484 S.W.3d 434 (Tex. Crim. App. 2016); Rosado v. State, 267 So.3d 4 (Fla. Dist. Ct. App. 2019) (“The notice stated that the Broward Sheriff’s Office Crime Laboratory inappropriately used Combined Probability of Inclusion (CPI) to calculate the statistical probabilities of genetic profiles in complex mixture DNA cases.”); Tex. Forensic Sci. Comm’n, Summary of Texas CPI Case Review Process, Sept. 15, 2016, <https://perma.cc/8MAS-EURJ>.

306. United States v. Carney, No. 1:20-cr-121, 2022 WL 1155902 (S.D. Ohio Apr. 19, 2022) (reasoning that manual deconvolution for three-person mixtures satisfies *Daubert* largely because it could be tested and was in widespread use and is still in use for some mixtures in 20% of laboratories); State v. Garland, 942 N.W.2d 732 (Minn. 2020); Roberts v. United States, 916 A.2d 922, 932 n.9 (D.C. 2007) (citing cases).

307. *But see supra* note 295 (on admissibility without a statistical assessment). The SWGDAM guidelines require a quantitative statement of some kind. SWGDAM, *supra* note 300, § 3.2.1 (“Except for a reasonably assumed contributor, the laboratory shall perform statistical analysis in support of any inclusion (or a ‘cannot be excluded’ conclusion) irrespective of the number of alleles detected and the quantitative value of the statistical analysis.”).

into two classes: inclusion–exclusion probabilities and likelihood ratios. The latter have greater support within the forensic DNA science community.³⁰⁸

Probability of inclusion or exclusion

Combined probability of inclusion (CPI) and exclusion (CPE). At first glance, it might seem that it is easy to describe the significance of a match to a mixed profile. We could simply ask what fraction of the population has genotypes that are not excluded by the alleles in the mixed profile. The answer in no way depends on the profiles of the victim, suspect, or person of interest; furthermore, no assumptions or inferences about the number of contributors are necessary for the calculation. For example, suppose we are sure that at one locus, the mixed profile indicates three alleles—A, B, and C, and no others—with population proportions 0.1, 0.2, and 0.3, respectively. The single-locus genotypes of people who could not be excluded as contributing to the profile are AA, BB, CC, AB, AC, and BC. Their (Hardy-Weinberg) probabilities are $0.1^2 + 0.2^2 + 0.3^2 + 2(0.1)(0.2) + 2(0.1)(0.3) + 2(0.2)(0.3) = 0.36$.³⁰⁹ So 36% of the population could not be excluded at this locus. Performing the same calculation at the other loci and multiplying them (linkage equilibrium) gives the “combined probability of inclusion,”³¹⁰ often referred to as CPI or RMNE (for “random man not excluded,” although its use is not restricted to male contributors). The combined probability of exclusion (CPE) is 1 minus the CPI.

Despite its computational simplicity,³¹¹ CPI has been deprecated as “hard to justify”³¹² and an “interim” measure.³¹³ In the three-allele mixture, for example,

308. See *infra* section titled “Likelihood ratios (LRs).”

309. See *supra* note 169.

310. Adjustments for population structure (see *supra* section titled “Could an Unrelated Person Be the Source?”) can be applied. Examples of how to compute CPI in different scenarios are given in SWGDAM, *supra* note 300, § 4C.

311. Determining the set of loci and alleles to use in the computation can be anything but simple. On the subtleties in using CPI properly, see *Barber v. United States*, 179 A.3d 883, 892 n.17 (D.C. 2018); Bieber et al., *supra* note 300; Coble et al., *supra* note 264, at 243–44; SWGDAM, *supra* note 300, § 4C.

312. NRC II, *supra* note 7, at 130.

313. Bieber et al., *supra* note 300, at 3. PCAST, *supra* note 105, at 82, expressed skepticism of the “foundational validity” of CPI, and stressed that “DNA analysis of complex mixtures should move rapidly to more appropriate methods based on probabilistic genotyping.” The group added that “[i]f, for a limited time, courts choose to admit results based on the application of CPI, validity as applied would require that, at a minimum, they be consistent with the rules specified in the paper.”

it counts pairs of people who could not jointly be the two contributors.³¹⁴ Thus, it does not actually resolve the mixture profile into the genotypes of those individuals who probably contributed to the mixed sample. Moreover, it disregards peak-height information, and it ignores the fact that there may be a known contributor such as the victim.

The number of reported opinions responding to challenges to CPI as lacking scientific validity or general acceptance is relatively small. Some courts have reasoned that the CPI is no different than the random-match probability accepted in single-source cases.³¹⁵ More often, courts recognize that the CPI has been criticized in scientific literature but conclude that it remains generally accepted as a reasonable or often conservative way to express the significance of an inclusion.³¹⁶

Random-match probability (RMP). Specifying the number of contributors in the mixture allows the expert to compute a somewhat different probability of inclusion, called the random-match probability (RMP). The RMP also can use peak heights and other information to deconvolve the mixed profile in light of any known or assumed contributor profiles. After taking all these things into account, the analyst will be left with a reduced set of plausible genotypes of the contributors. The RMP is just the sum of the probabilities for each of these genotypes. For example, suppose the victim's account and other information in a rape case establishes that one man committed the rape, and the victim's genotype at a locus is AB. DNA from a vaginal swab demonstrates the presence of three alleles A, B, and C with the peak for A and B being about the same height and C being much lower. The male contributor's genotype could be AC, BC, or CC,³¹⁷

314. For instance, the genotype pairs AA/AA, AA/BB, AA/CC, AA/AB, and AA/AC can be ruled out. They can produce no more than two peaks in the mixture electropherogram.

315. *E.g.*, *People v. Sandifer*, 65 N.E.3d 969, 980 (Ill. App. Ct. 2016); *cf.* *Coy v. Renico*, 414 F. Supp. 2d 744, 762 (E.D. Mich. 2006) (as “mere[] extensions of the generally accepted and generally reliable techniques applied to single source DNA samples,” CPI and CPE satisfy due process).

316. *E.g.*, *State v. Bander*, 208 P.3d 1242 (Wash. Ct. App. 2009); *cf.* *State v. Garland*, 942 N.W.2d 732, 744 (Minn. 2020) (relying on unspecified “evidence before the district court . . . including a peer-reviewed scientific article regarding best practices for using CPE” to conclude that “CPE has long been accepted by the forensic community as a valid method for calculating the probability of exclusion from a DNA mixture”).

317. The rapist must be responsible for the small peak C. Because it is small, the man is a minor contributor to the sample. A genotype of CC would explain the pattern of major peaks for A and B and a minor peak for C. Genotypes AC and BC for the rapist also would explain this pattern. For a small quantity of DNA of type AC, the A peak gets boosted slightly but remains approximately equal to the B peak. For a minor contribution of genotype BC, the B peak gets boosted, but only slightly. Because no other male genotypes (AA, AB, and BB) can explain the minor peak C in the mixture profile, the rapist must be genotype AC, BC, or CC, as stated in the text.

and the expert could estimate the frequency of these genotypes (using the population allele frequencies posited in the previous subsection) as $2(0.1)(0.3) + 2(0.2)(0.3) + (0.3)(0.3) = 0.27$. Instead of reporting the 36% for the single-locus inclusion probability as in the CPI calculation, the expert could report a smaller inclusion probability of 27%.³¹⁸

Testimony about the RMP (or the frequency with which unrelated individuals would match the deduced, smaller set of genotypes) is often encountered.³¹⁹ Sometimes the RMP or frequency is described as if it were the CPI.³²⁰ Issues likely to arise with the RMP are whether identification of major and minor alleles is correct and how the number of contributors was determined.

Likelihood ratio (LR)

Definition of the likelihood ratio. Rather than testify to inclusion probabilities for close relatives or for randomly selected, unrelated members of the suspect population, a DNA expert can present a likelihood ratio (LR) to express the evidentiary value of a DNA comparison.³²¹ In the most general formulation,

318. More examples and more refined calculations are in SWGDAM, *supra* note 300, § 4A; *see also* Bright & Coble, *supra* note 300, at 12–14. In two-person mixture cases in which it is reasonable to assume that the victim's DNA is present, the other contributor's genotype might be clear from the remaining peaks, and the RMP will be the same as in a single-source case for that contributor.

319. *E.g.*, United States v. Carney, No. 1:20-cr-121, 2022 WL 1155920 (S.D. Ohio Apr. 19, 2022) (RMP for the major contributor to a three-person mixture was said to be “1 in 299 septillion 400 sextillion” and therefore dispositive even though no alleles from the minor contributors could be reported); United States v. Graves, 465 F. Supp. 2d 450, 453 (E.D. Pa. 2006) (RMPs of 1/2, 1/2,900, and 1/3,600); People v. Roberts, 283 Cal. Rptr. 3d 357 (Cal. Ct. App. 2021); State v. Gonzalez, 204 A.3d 1183, 1189 (Conn. App. Ct. 2019); State v. Lowery, 427 P.3d 865 (Kan. 2018); State v. Hannon, 703 N.W.2d 498, 508–09 (Minn. 2005).

320. *E.g.*, People v. Brown, 82 N.E.3d 148, 169–70 (Ill. App. Ct. 2017) (The state's expert “testified that based on the only 6-loci analysis, he could determine only that defendant could not be excluded as a contributor to the major profile. . . . [T]he frequency which estimated the chance a random person would be included as a contributor in that major DNA profile [was] 1 in 670,000 black, 1 in 580,000 white, or 1 in 6.1 million Hispanic unrelated individuals. . . . The DNA Advisory Board has endorsed two methods . . . in cases of mixed DNA samples: (1) the combined probability of inclusion (or its reverse, the combined probability of exclusion) or (2) the likelihood ratio calculation.”). It is not always clear from an opinion whether the expert is giving a CPI or an RMP. *E.g.*, State v. Dwyer, 985 A.2d 469, 478–79 (Me. 2009) (giving “inclusion probabilities” without indicating whether they were CPIs or RMPs). As a result, some of the opinions cited above as illustrative of one method or the other could be misclassified.

321. In civil and criminal cases that involve DNA testing for parentage, LR's are routinely introduced under the name “paternity index.” 1 McCormick on Evidence, *supra* note 1, § 211.

an LR is the ratio between (1) the probability of the observed data when one hypothesis about the origin of the data is true and (2) the probability of the data when a non-overlapping hypothesis is true.³²² In symbols,

$$LR = \frac{P(\text{data} | H_1)}{P(\text{data} | H_2)}$$

where the vertical line stands for “given” or “conditional on.” For DNA evidence, H_1 could be a “contributor hypothesis” that names the defendant as the contributor or one of the contributors, and H_2 could be a “noncontributor hypothesis” that names individuals other than the defendant as the contributors. When multiple contributors are likely, various pairs of hypotheses may need to be considered to avoid misstating the value of the evidence.³²³

The baseline for understanding the value of the LR is 1—the ratio when the probabilities are equal. Equal evidence probabilities mean that the evidence is not probative of which hypothesis is true. For example, if H_1 asserts that a defendant’s DNA is in a trace-evidence sample, and H_2 asserts that the defendant’s identical twin’s DNA is in the sample, then $LR = 1$. The DNA data have the same probability under each scenario and therefore are not probative of which twin is a contributor to the sample.

The value of the LR normally would be different if the noncontributor hypothesis H_2 were that a different relative or an unrelated person is a contributor. Consider a single-source sample from the crime scene: If the STR profile of the sample is more probable if the defendant’s DNA is in it than if the alternative person’s DNA is there, then the LR exceeds 1. The profile data are more compatible with the defendant’s being the source (H_1) than with the alternative (H_2). As such, the data can be said to support H_1 (compared to H_2). Conversely, if the LR is less than 1, the data support H_2 .

In the legal literature, LRs often are presented as a way to quantify probative value.³²⁴ The idea is that relevant evidence changes the probability of some material fact, and the probative value of this evidence is the extent of the change. As noted in “Appendix: Conditional Probability and Bayes’ Rule” of the *Reference Guide on Statistics and Research Methods*, in this manual, under certain conditions, the change in the odds (known to statisticians as the Bayes’ factor) is simply the LR. When this is the case, the odds after receiving the evidence (the posterior odds) are just the odds before receiving the evidence times the LR:

$$\text{posterior odds} = LR \times \text{prior odds}$$

322. Mathematically, when the variable is essentially continuous (such as the height of a peak on an electropherogram) rather than discrete (such as the number of tandem sequences in an STR allele deduced from a peak), the LR is a ratio of probability densities. The definition above also omits a proportionality constant in both the numerator and denominator that cancels out when the ratio is formed.

323. Richard Wivell et al., *An Investigation into Compound Likelihood Ratios for Forensic DNA Mixtures*, 14 *Genes* 714 (2023), <https://doi.org/10.3390/genes14030714>.

324. See 1 McCormick on Evidence, *supra* note 1, § 185.

This version of Bayes' rule reinforces the idea that evidence with an LR of 1 has no probative value. Multiplying by 1 has no effect on the odds of H_1 relative to H_2 .

LRs are versatile. With DNA evidence, they can be used to convey the probative value of an inclusion for complex, mixed samples as well as for single-source samples with respect to a variety of alternative explanations for the composition of the sample. For instance, the hypothesis advanced by the prosecution might be that the “defendant and his deceased wife . . . were contributors to a DNA mixture profile drawn from a blood stain on defendant’s sweatshirt,” and the contrasting hypothesis could be that DNA from “two randomly selected individuals” was in the stain.³²⁵

Moreover, LR_s do not require a binary, include-exclude determination. They do not even require analytical and stochastic thresholds for allele determinations. As such, they can reflect more of the information in the electropherograms than both the CPI and the RMP can. Many forensic geneticists and statisticians regard them as particularly advantageous for mixture interpretation,³²⁶ and laboratories across the world have turned to probabilistic-genotyping software to produce LR_s to interpret electropherograms from low-template, degraded, or mixed DNA samples with greater consistency than unassisted human judgment can provide.³²⁷

In evaluating this development, it is important to keep in mind that the LR is merely a mathematical expression. The utility of any stated value for the quantity depends on how it was computed. Are the hypotheses or propositions appropriate to the case at hand? This is a question of relevance. Has the computational system been shown to give numbers that properly discriminate between the propositions for the DNA samples in the case? This is a question of validity. Assuming validity, how should LR_s be presented to lay factfinders so that they

325. *People v. Ramsaran*, 79 N.E.3d 1120 (N.Y. App. Div. 2017).

326. *E.g.*, Bright & Coble, *supra* note 300, at 11 (“The LR is our preferred approach to interpreting forensic DNA mixtures.”); Royal Soc’y, *Forensic DNA Analysis: A Primer for Courts* 36 (2017) (“Likelihood ratios are generally accepted as being the most appropriate method for evaluating the evidential strength of DNA profiles.”); Peter Gill et al., *DNA Commission of the International Society for Forensic Genetics: Assessing the Value of Forensic Biological Evidence—Guidelines Highlighting the Importance of Propositions: Part I: Evaluation of DNA Profiling Comparisons Given (Sub-) Source Propositions*, 36 *Forensic Sci. Int’l: Genetics* 189 (2018), <https://doi.org/10.1016/j.fsigen.2018.07.003>; Peter Gill et al., *DNA Commission of the International Society of Forensic Genetics: Recommendations on the Evaluation of STR Typing Results that May Include Drop-out and/or Drop-in Using Probabilistic Methods*, 6 *Forensic Sci. Int’l: Genetics* 679, 684 (2012), <https://doi.org/10.1016/j.fsigen.2012.06.002> (“probabilistic approaches and likelihood ratio principles are superior to classical methods”). United Kingdom guidelines regard statements of LR_s as the only acceptable approach to mixture interpretation. Forensic Science Regulator, Guidance: DNA Mixture Interpretation (FSR-G-222 Issue 3) (2020), <https://perma.cc/L34N-Y54X>.

327. For a concise history of the transition to probabilistic genotyping systems, see Michael D. Coble & Jo-Anne Bright, *Probabilistic Genotyping Software: An Overview*, 38 *Forensic Sci. Int’l: Genetics*, 219, 219–21 (2019), <https://doi.org/10.1016/j.fsigen.2018.11.009>.

fairly convey the weight of the evidence with respect to the relevant hypotheses? This is a question of probative value and prejudice. The remaining subsections elaborate on these points and describe the ways in which LR_s for DNA evidence are computed.

Binary genotyping. Traditional forensic STR-CE profiling is “binary” or “deterministic.” Like a baseball umpire calling a pitch a strike or a ball, an analyst (or software) labels a peak as either an allele or not. An allele call requires that the peak height exceed an analytical threshold and that the peak not look like an artifact.³²⁸ At each locus, the analyst decides whether there is one allele (homozygosity) or two different alleles (heterozygosity) on the pair of paternal and maternal chromosomes. If the reported profiles in two samples are consistent, there is a profile match. At that point, taking the reported matching profile as 100% certain, analysts can quote LR_s for inclusions as the reciprocal of the RMP.³²⁹ Thus, the data from the laboratory instruments are processed in two phases—a binary match/no-match phase, and a separate phase to give an LR for the categorical match (by regarding the analyst’s inclusion of the defendant as an adequate summary of the data).

This two-stage procedure works well enough in many cases (and in the absence of better alternatives). The 1996 NRC committee report endorsed then-available binary LR_s (based solely on the positions of DNA fragments subjected to electrophoresis) as opposed to the CPI.³³⁰ SWGDAM promulgated guidelines for calculating these LR_s.³³¹ Despite objections based on *Frye* or *Daubert*, appellate courts that have considered binary LR_s in mixture cases have upheld their admission.³³²

328. See *supra* section titled “Miniaturization for rapid capillary electrophoresis” n.52 (before Table 5); *supra* section titled “Did the Sample Contain Enough DNA?”

329. *E.g.*, *Augillard v. Madura*, 257 S.W.3d 494, 498–99 (Tex. Ct. App. 2008) (“the test showed a complete match at all seventeen DNA markers with a likelihood ratio exceeding one trillion”). The LR is the reciprocal of the RMP if the match, as determined by thresholds and human judgment, has probability 1 when H_1 is true (if the defendant is a source, the analyst is certain to report an inclusion) and probability RMP when H_2 is true (if a random, unrelated person is a source, the probability of inclusion is just the random-match probability).

330. NRC II, *supra* note 7, at 129 (“when the contributors to a mixture are not known or cannot otherwise be distinguished, a likelihood-ratio approach offers a clear advantage and is particularly suitable”).

331. SWGDAM, *supra* note 300, § 4(B).

332. *E.g.*, *State v. Garcia*, 3 P.3d 999 (Ariz. Ct. App. 1999) (applying *Frye*); *Commonwealth v. Gaynor*, 820 N.E.2d 233, 252 (Mass. 2005); *People v. Coy*, 669 N.W.2d 831, 835–39 (Mich. Ct. App. 2003) (incorrectly treating mixed-sample likelihood ratios as a part of the statistics on single-source DNA matches that had already been held to be generally accepted); *cf.* *Coy v. Renico*, 414 F. Supp. 2d 744, 762–63 (E.D. Mich. 2006) (likelihood ratio for a mixed stain in *People v. Coy*, *supra*, was consistent with due process).

Probabilistic genotyping software (PGS). Binary LR_s have significant limitations. Using only the genotypes as reported from the initial match/no-match phase overlooks any uncertainty in the allele calls and related single-locus genotype determinations. In particular, it ignores the possibility of allelic drop-out or drop-in, which could make the assignment of possible genotypes less than certain. It tosses out any alleles that might be producing peak heights below the laboratory's analytical threshold (or that are wrongly thought to be artifacts).³³³ In effect, a peak above the line is treated as if it has probability 1 for being an allele, while a peak just below the line is treated as if it has probability zero of being an allele. But surely, if an allele is present (or absent) in a sample, then the probability of a peak just below the threshold is not radically different from a peak just above the line. Newer, probabilistic genotyping systems overcome these limitations by computing a likelihood ratio $P(\text{data}|H_1) \div P(\text{data}|H_2)$ without having to declare a match or inclusion.

They do so by using weighted averages, for each hypothesis, of the probabilities that all sets of genotypes that could be in the complicated mixture profile are in fact there—based on how often the possible genotypes occur in the relevant population.³³⁴ These “prior probabilities” are weighted by the probability that the genotype set (if present) would give rise to the pattern in the electropherogram.³³⁵ Thus, a possible genotype set that is more likely to give rise to the electropherogram if it is assumed to be in the mixture profile gets more weight.³³⁶

Over the past two decades, researchers have developed a variety of PGS programs with increasingly complete statistical models of the production of true peaks and artifacts in electropherograms. The simplest programs (called qualitative, discrete, or semi-continuous) consider only the alleles that might be

333. See, e.g., Butler, *supra* note 113, at 39–40.

334. For example, under a contributor hypothesis H_1 , such as “the defendant and a brother are the source of the semen,” there might be just one set of genotypes at a given locus that conceivably could have produced the electropherogram. For concreteness, let's say this set is AB/AC, meaning that one brother is AB and the other is AC. Under a noncontributor hypothesis H_2 of two random, unrelated men, there might be two sets, such as AB/AC and AA/BC. The prior probability of each of these genotype sets comes from population-genetic models and allele-frequency databases.

335. The genotypes-to-data probability, $P(\text{data}|\text{genotypes})$, depends entirely on the sample and the CE measurement process. It is the same under the contributor and the noncontributor hypotheses.

336. The mathematical expression for the weighted averaging is in Peter Gill et al., *A Review of Probabilistic Genotyping Systems: EuroForMix, DNASTatX and STRmix™*, 12 *Genes* 1559, at 2 (2021), <https://doi.org/10.3390/genes12101559>, and Mark W. Perlin & Alexander Sinelnikov, *An Information Gap in DNA Evidence Interpretation*, 4 *PLoS ONE* e8327, at 2 (2009), <https://perma.cc/RZ3T-KLHV>.

present³³⁷ and do not model the height of the peaks.³³⁸ They incorporate estimated probabilities of allelic drop-in and drop-out³³⁹ when assigning weights on the prior probabilities.³⁴⁰ The output is an LR that draws on more of the information in the mixture electropherogram.³⁴¹ One such program, the Forensic Statistical Tool developed at the New York City Office of the Chief Medical Examiner, was the subject of considerable litigation³⁴² and is no longer used.³⁴³

These programs have been supplanted by “continuous” or “quantitative” systems that model peak heights themselves, automatically accounting for drop-in and drop-out as well as stutter peaks, degradation, and other complicating factors. Two types of continuous systems have been developed. One type posits an explicit mathematical function relating to peak heights. It uses a common statistical method for estimating parameters and applies the estimates to find the genotypes-to-data probabilities.³⁴⁴ The other type relies on Bayes’ rule to obtain these weights.³⁴⁵

337. Even in this regard, many “semi-continuous methods do not model artefacts such as stutter. In these models, peaks must be assigned as stutter or labelled as allelic by an analyst prior to interpretation [by software].” Jo-Anne Bright et al., *A Series of Recommended Tests When Validating Probabilistic DNA Profile Interpretation Software*, 14 *Forensic Sci. Int’l: Genetics* 125, 126 (2015), <https://doi.org/10.1016/j.fsigen.2014.09.019>.

338. At most, they use peak heights in more limited ways, “to inform the probability of drop-out or to infer a major donor genotype by applying different drop-out probabilities per contributor.” Nevertheless, “[t]hese systems do represent an advance over the binary model as they can take account of multiple contributors, low-template DNA and replicated samples.” Gill et al., *supra* note 336, at 3.

339. Drop-in and drop-out probabilities come from experiments on how often peaks from samples with known alleles disappear (drop below the analytical threshold) or appear when the known samples do not contain known alleles for those peaks. See Coble & Bright, *supra* note 327, at 221.

340. For a numerical example of how drop-in and drop-out probabilities are incorporated, see Michael Coble et al., *Mixtures and Probabilistic Genotyping*, in *Forensic DNA Applications: An Interdisciplinary Perspective* 155, 164–66 (Dragan Primorac & Moses Schanfield eds., 2d ed. 2023).

341. For details on the operation of a leading qualitative system, see Gill et al., *supra* note 36, at 129–80 (on LRmix and LRmix Studio); see also Bright & Coble, *supra* note 300, at 107–40.

342. Compare, e.g., *United States v. Jones*, 965 F.3d 149, 162 (2d Cir. 2020) (“no abuse of discretion in the district court’s conclusion [of] reliability sufficient to support admission” of an LR likelihood ratio of “1,340, showing very strong support”), with *State v. Rochat*, 269 A.3d 1177 (N.J. App. Div. 2022) (FST not generally accepted).

343. *Jones*, 965 F.2d at 152.

344. On the general nature of parameters in a statistical model and their estimation, see David H. Kaye & Hal S. Stern, *Reference Guide on Statistics and Research Methods*, in this manual. In one widely used program, EuroForMix, the statistical model for differences between estimated and expected peak heights is known as the gamma (γ) distribution. Gill et al., *supra* note 36, at 199–200; Gill et al., *supra* note 336, at 13–14. The parameters of this distribution are closely related to the proportion of mixture DNA from each contributor and the coefficient of variation of peak heights across all the loci. Gill et al., *supra* note 36, at 201–05. The parameter values are estimated by the maximum-likelihood method. *Id.* at 461.

345. Gill et al., *supra* note 36, at 462. Because peak height is a continuous variable, the version of Bayes’ rule for discrete variables presented in the *Reference Guide on Statistics and Research Methods*,

The two programs that are most commonly used in the United States are of the latter type. They have the brand names TrueAllele and STRmix. To approximate the genotypes-to-data probabilities, they take a kind of random walk through the space of all possible parameter values,³⁴⁶ spending more time at those combinations at which the pattern of peaks computed by the model come closest to the observed values.³⁴⁷ The relative time spent at each such combination determines the weights $P(\text{data}|\text{genotype set})$.³⁴⁸ The technical name for this probabilistic approximation method is Markov chain Monte Carlo (MCMC).³⁴⁹ The “Monte Carlo” part refers to computer simulation of the outcomes of any process that has some randomness in it.³⁵⁰ Simulations rarely give exactly

in this manual, needs to be restated with a probability density function and integration instead of summation. The underlying concepts are the same, however, and, for ease of presentation, this reference guide has not distinguished between discrete probability distributions and probability density functions. The LR for continuous models is a ratio of probability densities.

346. For partial descriptions of the parameters in the peak-height model of STRmix, see Bright & Coble, *supra* note 300, at 143 (“For the simplest profile . . . these are 1. DNA amount (template) for each contributor to the profile 2. The degradation rate for each contributor to the profile 3. Amplification efficiencies for each locus within the profile.”); Coble et al., *supra* note 340, at 167 (“template quantity, degradation, stutter ratios, etc.”).

347. At least, that is the expectation. The game of “hot and cold,” in which a person searching for an object in a room is told whether he is getting closer to or farther from the object as he moves around the room is sometimes offered as a metaphor for the computer-simulation procedure. *E.g.*, *United States v. Gissantener*, 417 F. Supp. 3d 857 (W.D. Mich. 2019), *rev'd*, 990 F.3d 457 (6th Cir. 2021); *United States v. Lewis*, 442 F. Supp. 3d 1122, 1142 (D. Minn. 2020) (Rep. & Recommendation) (extending the analogy to multiple rooms). A better (but still imperfect) analogy might be trying to climb the highest mountain in a region with no visibility by taking proposed steps that increase the elevation with a higher probability than steps that do not, until no proposed step increases the altitude. This should get the climber to the summit if the mountain is smooth and if there are no smaller mountains in the region. Unlike the hot-and-cold game in which one *knows* when the object is found, in the latter landscape, there is a risk of being fooled into thinking that a smaller peak is the highest there is.

348. See Coble et al., *supra* note 340, at 167–68.

349. The “Markov chain” part refers to situations in which the next value of a random variable depends to some degree on just the previous value (like a peculiar coin that is a little more likely to turn up heads when tossed if a head had turned up on the previous toss). “Since their invention [in 1905], Markov chains have become extremely important in a huge number of fields such as biology, game theory, finance, machine learning, and statistical physics.” Joseph K. Blitzstein & Jessica Hwang, *Introduction to Probability* 459 (2015). Nonetheless, the success of an MCMC algorithm in one domain or type of problem does not guarantee that it will work as well in another situation.

350. For example, if we wanted to estimate the probability of a lucky streak of ten heads when tossing a fair coin, we could program a computer to pick a 0 or a 1 at random ten times and see if the sequence was 1111111111. One such simulation would not tell us much, but if the computer simulated ten tosses one million times, keeping track of how many trial sequences were 1111111111, the proportion usually would be close to the true value of the probability. That the inventors of the computer-simulation method referred to a famous gaming casino in naming it is an historical accident. See Nicholas Metropolis, *The Beginning of the Monte Carlo Method*, Los Alamos Sci., Special Issue 125, 127 (1987), available at <https://perma.cc/M53S-Y9X9>.

identical estimates if they are rerun, and “[a] simulation result is easy to criticize: how do you know you ran it long enough? How do you know your result is close to the truth? How close is ‘close?’”³⁵¹

These are questions to be addressed in validation studies, considered in the next subsection. Proof of validity also needs to attend to the more fundamental issue of the adequacy of the parametric models for predicting the observed data from the possible genotype sets. PGS programs that use Bayesian methods and MCMC computations may not use exactly the same statistical models, and one program may require more preliminary input from the DNA analyst or laboratory using it than another.³⁵² Rather than a blind acceptance of the output, the reasonableness and impact of these choices may need to be examined.

Validation of PGS programs. In the courtroom, the most important output of PGS is the LR for a set of hypotheses that include the defendant as a contributor to the mixture (or to a single-source sample, typically of very low quantity).³⁵³ The LR is not a measured physical or chemical property, like height or weight. Rather, it is an expression of the degree to which the electropherogram is more indicative of scenarios that include the defendant as a contributor than it is of scenarios that do not. How, then, do scientists who have written the software convince other scientists that the values of the LR are believable?

Validation has at least three components. First, the mathematical equations used in the software and the ways of solving those equations must be appropriate.³⁵⁴

351. Blitzstein & Hwang, *supra* note 349, at 421 (also noting that “[t]he simulations may need to run for a vast, unknown amount of time to get decent answers”).

352. See *Daniels v. State*, 312 So.3d 926 (Fla. Dist. Ct. App. 2021) (whereas STRmix requires laboratory-specific values from “internal validation” studies, TrueAllele does not). TrueAllele also dispenses with analytic thresholds; but STRmix uses them. William C. Thompson, *Uncertainty in Probabilistic Genotyping of Low Template DNA: A Case Study Comparing STRMix™ and TrueAllele™*, J. Forensic Sci. (2023), <https://doi.org/10.1111/1556-4029.15225>. However, STRmix may be moving away from these rigid cutoffs. See Duncan Taylor & John Buckleton, *Combining Artificial Neural Network Classification with Fully Continuous Probabilistic Genotyping to Remove the Need for an Analytical Threshold and Electropherogram Reading*, 62 Forensic Sci. Int'l: Genetics 102787 (2023), <https://doi.org/10.1016/j.fsigen.2022.102787>.

353. Some PGS programs are also useful for estimating the number of contributors and mixture ratios and for producing a list of high-LR contributor genotypes that could be submitted to an offender DNA database.

354. The conceptual underpinnings of the two main continuous PGS systems have not attracted major criticism. The weighted averaging described in the section titled “Probabilistic genotyping software (PGS)” comes out of the basic mathematics of probability theory and Bayes’ rule. The values for the prior probabilities are population genotype frequencies that are calculated from population-genetics models and are discussed in the section titled “Inference, Statistics, and Population Genetics in Human Nuclear DNA Testing” above. The modeling of the genotypes-to-data probabilities also

In litigation, the general mathematical structure behind most PGS software has not been seriously questioned.³⁵⁵

Second, the coding of the mathematical formulas and techniques for solving the equations must be correct. Examining parts of the source code of a program and testing whether modules work as they are supposed to as the program is written and revised can reveal coding and logical problems.³⁵⁶ Thus, voluntary quality-assurance standards for software (and hardware) “verification and validation” are well known among software engineers.³⁵⁷ These standards are sometimes cited to support motions for discovery of all of the source code of a PGS program or to buttress an argument that the program cannot be found to be scientifically valid or generally accepted without publishing the code (open-access software), or at least affording defense experts extensive review of proprietary software, perhaps under a confidentiality order.³⁵⁸

needs to be sound. Finally, the statistical procedures (like an MCMC algorithm, which is not itself the genetic or statistical model at the heart of PGS) must be capable of solving (at least approximately) the equations to which they are applied.

355. *E.g.*, *United States v. Lewis*, 442 F. Supp. 3d 1122, 1145 (D. Minn. 2020) (Rep. & Recommendation) (“The underlying principles on which STRmix is based—the MCMC, the Hastings-Metropolis algorithm, and Bayesian statistics—are not themselves challenged as unreliable. Nor is the purely mathematic calculation of relative probabilities (i.e., the likelihood ratio).”); *People v. Davis*, 290 Cal. Rptr. 3d 661, 679 (Cal. Ct. App. 2022) (“[t]he scientific and mathematical principles behind STRmix are well-established and widely-accepted in the scientific community”); *People v. Wakefield*, 195 N.E.3d 19, 28 (N.Y. 2022) (“It was undisputed that the foundational mathematical principles [of TrueAllele] (MCMC and Bayes’ theorem) are widely accepted in the scientific community.”). *But see* *United States v. Gissantaner*, 417 F. Supp. 3d 857, 870 (W.D. Mich. 2019), *rev’d*, 990 F.3d 457 (6th Cir. 2021) (suggesting that software that implements Bayes’ rule is problematic because “Bayesian reasoning ‘breaks down in situations where information must be conveyed from one person to another such as in courtroom testimony’”).

356. “Black box” testing, which involves no inspection of the underlying code, is also a standard procedure. *See, e.g.*, Christopher Henard et al., *Comparing White-Box and Black-Box Test Prioritization*, ICSE ’16: Proceedings of the 38th International Conference on Software Engineering 523 (2016), <https://doi.org/10.1145/2884781.2884791>.

357. In this context, “verification” refers to making sure that the software does what the specifications call for. “Validation” refers to ensuring that the product meets the needs of its users. The ISO/IEC/IEEE 29119 series of standards offer guidance on “V & V” for business and other environments. *See* ISO/IEC/IEEE 29119–1:2022(en) (Software and systems engineering—Software testing—Part 1: General concepts). They go beyond establishing scientific or technical merit.

358. *See, e.g.*, *People v. Wakefield*, 195 N.E.3d 19 (N.Y. 2022); *State v. Watkins*, 648 S.W.3d 235 (Tenn. Ct. Crim. App. 2021); *cf.* Nathaniel Adams et al., *Letter to the Editor—Appropriate Standards for Verification and Validation of Probabilistic Genotyping Systems*, 63 J. Forensic Sci. 339 (2018), <https://doi.org/10.1111/1556-4029.13687> (urging developers to utilize the IEEE V & V standards). Without necessarily relying on the V & V software engineering standards, legal academics tend to advocate disclosure of source code. *See* Rebecca Wexler, *Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System*, 70 Stan. L. Rev. 1343, 1376 (2018) (citing articles). The authors of some PGS programs are more skeptical of the need for defendants’ experts or third parties to peruse the code itself. *E.g.*, Duncan A. Taylor et al., *Commentary: A “Source” of*

The third component of validation is direct testing of the PGS by presenting “a wide range of evidence types, representing a typical case, as well as extreme examples.”³⁵⁹ The performance on data from a large number of mixtures of known composition can provide convincing empirical proof of the validity of a conceptually sound system (for mixtures like those in the experiment). Mixed samples can be constructed in the laboratory by combining DNA in varying ratios and total amounts from known individuals.³⁶⁰ Mixed profiles for validation also can be simulated by combining known single-sample-profile data together mathematically.³⁶¹ Yet a third approach uses samples from cases in which defendants were convicted.³⁶² The idea in each approach is to assemble two sets of paired data for testing the PGS: contributor-mixture pairs and

Error: Computer Code, Criminal Defendants, and the Constitution, 8 *Frontiers Genetics* 33 (2017), <https://doi.org/10.3389/fgene.2017.00033> (maintaining that it is “nearly impossible to identify subtle errors in code by viewing the code”).

A different argument about PGS source code is that studying it is somehow like cross-examining an “artificial intelligence,” and the Confrontation Clause therefore gives the defendant in a criminal case the right to see the code itself when the AI’s output is testimonial under *Crawford v. Washington*, 541 U.S. 36 (2004). *Wakefield*, 195 N.E.3d at 44–45 (concurring opinion). A more straightforward constitutional analysis would look to the Due Process Clause and ask whether the need to inspect the source code is critical in light of other methods of testing the operation of the software.

359. Gill et al., *supra* note 336, at 10.

360. Samples also can be made from DNA with different degrees of degradation or other complicating factors.

361. Artificially generating the validation profiles rather than profiling “made-up samples” has been advocated to “control the input variables exactly and produce profiles where the expected answer is known. [I]t is effectively impossible to create an exact 1:1 [or other mixture-ratio] mixture in vitro, [whereas artificially mixed profiles] are simply tidy single source, major, minor and balanced profiles.” Bright et al., *supra* note 337, at 126.

362. Treating these defendants as true contributors gives us a set of presumed-contributor-mixture pairs for testing the PGS. A much larger set of known noncontributor-mixture pairs for testing can be created by substituting genotypes of the defendants in all the other (unrelated) cases for the actual defendant. For a study employing this design, see Mark W. Perlin et al., *New York State TrueAllele® Casework Validation Study*, 58 *J. Forensic Sci.* 1458, 1469 (2013), <https://doi.org/10.1111/1556-4029.12223>; cf. David H. Kaye et al., *Validating the Probability of Paternity*, 31 *Transfusion* 823 (1991), <https://doi.org/10.1046/j.1537-2995.1991.31992094670.x> (examining the empirical distribution of the likelihood ratio for HLA-typing results in civil paternity cases). One might object that the presumed-contributor-mixture pairs could include some falsely convicted defendants. Obviously, it would be better to restrict the true-pair group to only true (contributor-mixture) pairs, as can be done with manufactured mixtures. However, a fraction of noncontributor-mixture pairs in the presumed-contributor-mixture pairs can only make it harder to find that a valid PGS typically generates large LR_s for the group designated as true pairs. A clear difference in PGS performance, as between the two large groups of pairs, is thus still good evidence of validity. A failure to find a difference is more ambiguous. It could be explained as either a lack of validity or as a consequence of having too many falsely convicted offenders’ profiles in the true-pairs group.

noncontributor-mixture pairs.³⁶³ Each method of forming the two sets has advantages and disadvantages, and a series of studies with the different designs is best.

The PGS can be applied to these validation test sets in several ways. Analysts can verify that the program gives large LR_s for true pairs in simple cases (ones they can readily resolve without the software).³⁶⁴ As the information available from the mixture declines (because of more complex peaks and low-template samples), the LR values should drop as well. Beyond such rudimentary checks, rates of misleading LR_s can be determined.³⁶⁵ The observed proportion of cases in the validation test set for which the PGS LR points in the wrong direction (toward exclusion for contributors or toward inclusion for noncontributors)³⁶⁶ is a kind of “error rate” for the PGS.³⁶⁷

Nonetheless, these misleading-LR rates do not tell the whole story. Even if the overall rates of misleading evidence are small, how do we know that an LR of 1 million is dramatically more powerful than an LR of, say, 1,000 or 100?³⁶⁸ To address this question, validation studies can examine whether larger LR_s are indeed associated with smaller rates of misleading evidence. Do LR_s of 1,000 or

363. The false-pair cases can be formed by pairing the data for each mixture with many randomly generated genotypes (using the allele frequencies in any population of interest).

364. Bright et al., *supra* note 337, at 126.

365. See Richard Royall, *On the Probability of Observing Misleading Statistical Evidence*, 95 J. Am. Stat. Ass’n 760 (2000).

366. The tendency of the computed LR_s to exceed 1 for mixtures that include the defendant’s DNA has been called “sensitivity.” The tendency to be less than 1 for mixtures that do include the defendant has been called “specificity.” E.g., Scientific Working Group on DNA Analysis Methods (SWGDM), Guidelines for the Validation of Probabilistic Genotyping Systems § 21 (2015), accessible via <https://perma.cc/5V7F-6VJ7>. The terms are an extension of their more established meaning for binary tests. See *supra* section titled “Validity.”

367. *United States v. Gissantaner*, 990 F.3d 457, 465 (6th Cir. 2021); Colin Aitken & Franco Taroni, *The History of Forensic Inference and Statistics: A Thematic Perspective*, in *Handbook of Forensic Statistics* 3, 23 (David Banks et al. eds., 2021); David H. Kaye, *Forensic Statistics in the Courtroom*, in *id.* at 225. Other measures of performance also are in use. See Aitken & Taroni, *supra*, at 23–24; Daniel Ramos et al., *Validation of Forensic Automatic Likelihood Methods*, in *Handbook of Forensic Statistics* 143 (David Banks et al. eds., 2021). Repeatability of computed LR_s (also referred to as “precision” and “reproducibility” in the standards and writing on PGS) can be studied by rerunning the PGS on the same data. It is an issue for MCMC computations. See, e.g., David W. Bauer et al., *Validating True Allele Interpretation of DNA Mixtures Containing up to Ten Unknown Contributors*, 65 J. Forensic Sci. 380 (2020), <https://doi.org/10.1111/1556-4029.14204>.

368. The court in *United States v. Lewis*, 442 F. Supp. 3d 1122 (D. Minn. 2020), suggested that proof that “[t]he variation in the precise LR numbers generated by STRmix from one run to the next is very small” (deviating by no more than a factor of ten) answered this question. *Id.* at 1130. But replicability must not be mistaken for validity. Generating similar results on repeated tries is not necessarily the same as generating correct results. A scale that always reported objects with true weights of 1 gram and 1 kilogram as weighing 1–10 grams (first object) and 100–1,000 kilograms (second object) would massively overstate the weight of the heavier object.

more crop up in the noncontributor–mixture cases more than one time in 1,000? Do LR_s of 1 million or more arise more often than one time in a million in these cases? If not, the PGS LR_s are well calibrated (or at least not systematically overstated).³⁶⁹

The point at which enough well-designed studies have been conducted to answer these questions is not easily defined. Simply counting the number of studies with the word “validity” in their title only scratches the surface. Did the studies include large numbers of contributor and noncontributor pairs? Which performance metrics were used? Is the case within the range of conditions covered in the studies? In 2016, the President’s Council of Advisors on Science and Technology reported that PGS programs “clearly represent a major improvement over purely subjective interpretation,”³⁷⁰ but that the studies to that point, which came from the software developers rather than from completely separate researchers,³⁷¹ “have adequately explored only a limited range of mixture types (with respect to number of contributors, ratio of minor contributors, and total amount of DNA).”³⁷² Additional studies of as many as 10–person mixtures followed.³⁷³ Almost invariably, courts have upheld the admissibility of

369. Duncan Taylor et al., *Testing Likelihood Ratios Produced from Complex DNA Profiles*, 16 *Forensic Sci. Int’l: Genetics* 165 (2015), <https://doi.org/10.1016/j.fsigen.2015.01.008>. Simulating enough noncontributor pairs to test LR_s of very large magnitudes can be challenging. For one strategy, see Duncan Taylor et al., *Importance Sampling Allows H_d True Tests of Highly Discriminating DNA Profiles*, 27 *Forensic Sci. Int’l: Genetics* 74 (2017), <https://doi.org/10.1016/j.fsigen.2016.12.004>.

370. PCAST, *supra* note 105, at 79.

371. The report evinced concern over the involvement of the developers of the software in the validity studies of their products: “Appropriate evaluation of the proposed methods should consist of studies by multiple groups, *not associated with the software developers*, that investigate the performance and define the limitations of programs by testing them on a wide range of mixtures with different properties.” *Id.* (emphasis in original). The issue of developer involvement is discussed in *Gissantaner*, 990 F.3d 457 (6th Cir. 2021); *United States v. Lewis*, 442 F. Supp. 3d 1122, 1147–49 (D. Minn. 2020) (Rep. & Recommendation); *People v. Wakefield*, 195 N.E.3d 19 (N.Y. 2022); *People v. Williams*, 147 N.E.3d 1131 (N.Y. 2020); and other cases.

372. PCAST, *supra* note 105, at 80. The report added that “[t]he two most widely used methods (STRMix and TrueAllele) appear to be reliable within a certain range, based on the available evidence and the inherent difficulty of the problem. Specifically, these methods appear to be reliable for three–person mixtures in which the minor contributor constitutes at least 20 percent of the intact DNA in the mixture and in which the DNA amount exceeds the minimum level required for the method.” *Id.*

373. E.g., Michael S. Adamowicz et al., *Internal Validation of MaSTR™ Probabilistic Genotyping Software for the Interpretation of 2–5 Person Mixed DNA Profiles*, 13 *Genes* 1429 (2022), <https://doi.org/10.3390/genes13081429>; Bauer et al., *supra* note 367 (TrueAllele); Jo-Anne Bright et al., *Internal Validation of STRmix™—A Multi-Laboratory Response to PCAST*, 34 *Forensic Sci. Int’l: Genetics* 11 (2018), <https://doi.org/10.1016/j.fsigen.2018.01.003>; Sara Riman et al., *Examining Performance and Likelihood Ratios for Two Likelihood Ratio Systems Using the PROVEDIt Dataset*, 16 *PLoS ONE* e0256714 (2021), <https://doi.org/10.1371/journal.pone.0256714> (EuroForMix and STRmix applied

PGS results in the face of objections to their scientific validity or general scientific acceptance,³⁷⁴ although some cases have been remanded for more complete hearings.³⁷⁵

Presentation of LR. Whether likelihood ratios result from categorical matching or from PGS programs, courts may need to consider the manner in which the ratios are presented at trial. The legal concerns are avoiding overstating what science has to offer (an aspect of Rules 403 and 702) and preventing unfair prejudice or confusion (Rule 403). Like random-match probabilities, LR_s are easily transposed,³⁷⁶ and it can be argued that large numbers might be overvalued.³⁷⁷ With random-match probabilities, we saw that courts have reasoned that the possibility of transposition does not justify a blanket rule of exclusion.³⁷⁸ Appellate courts have not addressed the issue for LR_s,³⁷⁹ but it is not obvious

to the same mixture profiles). But in 2021, a small group of scientists and other personnel working at NIST released a draft of a “scientific foundation review” that stated “[c]urrently, there is not enough publicly available data to enable an external and independent assessment of the degree of reliability of DNA mixture interpretation practices, including the use of probabilistic genotyping software (PGS) systems.” Butler et al., *supra* note 105, at 6, 75.

374. *E.g.*, *Gissantaner*, 990 F.3d 457 (STRmix); *Lewis*, 442 F. Supp. 3d 1122 (STRmix); *People v. Davis*, 290 Cal. Rptr. 3d 661 (Cal. Ct. App. 2022) (STRmix); *State v. Simmer*, 935 N.W.2d 167 (Neb. 2019) (TrueAllele); *Wakefield*, 195 N.E.3d 19 (TrueAllele); *cf.* *United States v. Williams*, 382 F. Supp. 3d 928 (N.D. Cal. 2019) (LR inadmissible because “Bullet” PGS program was only validated for up to four-person mixtures, and it could not be shown that the mixture in question did not come from five individuals).

375. *E.g.*, *People v. Wortham*, 180 N.E.3d 516 (N.Y. 2021) (New York City Office of the Chief Medical Examiner’s program, FST).

376. For a few examples, see *People v. Pike*, 53 N.E.3d 147, 167 (Ill. App. Ct. 2016) (“estimates how much more likely it is that the suspect is the source of the evidence than it is that the evidence originated from a randomly selected member of the population unrelated to the suspect”); *State v. Pickett*, 246 A.3d 279 (N.J. App. 2021) (“a statistic measuring the probability that a given individual was a contributor to the sample against the probability that another, unrelated individual was the contributor”); *Commonwealth v. Treiber*, 121 A.3d 435, 445 (Pa. 2015) (“the hair was 1,000 times more likely to have come from appellant’s dog than any other dog”). These statements mischaracterize the likelihood ratio as the relative probability of a source hypothesis given the DNA data rather than the relative probability of the data given the hypotheses. Kaye et al., *supra* note 193, § 14.2.2.

377. Some commentary maintains that likelihood ratios are inherently prejudicial. *E.g.*, William C. Thompson, *DNA Evidence in the O.J. Simpson Trial*, 67 U. Colo. L. Rev. 827, 855–56 (1996); see also Richard C. Lewontin, *Population Genetic Issues in the Forensic Use of DNA*, in 1 *Modern Scientific Evidence: The Law and Science of Expert Testimony* § 17–5.0, at 703–05 (David L. Faigman et al. eds., 1st ed. 1998).

378. See section titled “Frequencies, Probabilities, and Prejudice” above.

379. Trial court rulings admitting likelihood ratios over the unfair prejudice objection were upheld in *State v. Ayers*, 68 P.3d 768, 778 (Mont. 2003), and *State v. Watkins*, 648 S.W.3d 235, 265 (Tenn. Crim. App. 2021).

why the rule would be different.³⁸⁰ The usual expectation is that “[a] district court concerned that the jury might misunderstand what the likelihood ratio means could require advocates to describe it in a way that will not generate unfair prejudice or mislead the jury.”³⁸¹

But what would that be? It has been suggested that experts should cease characterizing LRs as statements that “a match is X times more/less probable than a coincidence,”³⁸² for that formulation seems to invite the misunderstanding that the LR is the odds in favor of the contributor hypothesis. Better explanations of the LR would clarify that the value pertains to the probability of the evidence (the STR data) rather than the probability of the conclusion (who the contributors are). An LR of x means that the DNA data are x times more likely if the hypothesis that includes the defendant as a contributor is correct than if the hypothesis that only other people are contributors is correct. One textbook for practitioners advises that “[a]n example of suitable phrasing (where the LR = one billion) is ‘The evidence is a billion times more likely if the person of interest is a contributor than if a random, unrelated person is a contributor.’”³⁸³

Even with the correct conditional phrasing, however, the number may be puzzling. Factfinders are not used to thinking about ratios of probabilities. To help, experts could offer analogies. One such analogy is a partial fingerprint pattern that one in a thousand people would possess. A DNA comparison with

380. Psychological research to ascertain the effects of likelihood-ratio statements, as opposed to other quantitative and qualitative indicators of probative value, does not demonstrate that research subjects overvalue the evidence as described in questionnaires. See Busey, *supra* note 186; Edward K. Cheng, *The Burden of Proof and the Presentation of Forensic Results*, 130 Harv. L. Rev. F. 154, 161 (2017) (“An increasingly complex literature has emerged on lay understanding of likelihood ratios and how such quantitative information is best presented. Research thus far has yielded no easy answers”); Heidi Eldridge, *Juror Comprehension of Forensic Expert Testimony: A Literature Review and Gap Analysis*, 1 Forensic Sci. Int’l: Synergy 24 (2019), <https://doi.org/10.1016/j.fsisy.2019.03.001>; Martire, *supra* note 186. For some studies in this area, see Brandon L. Garrett et al., *Error Rates, Likelihood Ratios, and Jury Evaluation of Forensic Evidence*, 65 J. Forensic Sci. 1199 (2020), <https://doi.org/10.1111/1556-4029.14323>; William C. Thompson et al., *Perceived Strength of Forensic Scientists’ Reporting Statements About Source Conclusions*, 17 L. Probability & Risk 133 (2018), <https://doi.org/10.1093/lpr.mgy012>; William C. Thompson & Eryn J. Newman, *Lay Understanding of Forensic Statistics: Evaluation of Random Match Probabilities, Likelihood Ratios, and Verbal Equivalents*, 39 L. & Hum. Behav. 332 (2015), <https://doi.org/10.1037/lhb0000134>.

381. United States v. Gissantner, 990 F.3d 457, 470 (6th Cir. 2021) (internal quotation marks and alteration omitted).

382. Thompson, *supra* note 352; cf. Eur. Network of Forensic Sci. Inst., Best Practice Manual for Human Forensic Biology and DNA Profiling 29 (ENFSI-DNA-BPM-03, Ver. 01. 2022) (“Caution is required when using the word ‘match’ in statements because it might imply ‘identity’. The expert avoids any verbal statement that might imply that he/she is making an opinion on the identity of the questioned DNA.”).

383. Bright & Coble, *supra* note 300, at 30.

an LR of 1,000 has the same probative value as the partial print even though the phenomena have nothing to do with each other.³⁸⁴ Experts also could describe an LR as a measure of support for a scenario and add a qualitative expression such as “weak,” “strong,” “very strong,” and so on for the degree of support.³⁸⁵ Forensic statisticians originally proposed “verbal scales” for less objectively determined LRs for all kinds of forensic-science identification evidence, from fingerprints to toolmarks to handwriting, in the hope that these scales would encourage greater standardization in expressing the strength of the evidence and better comprehension of LRs.³⁸⁶ The Department of Justice’s Uniform Language for Testimony and Reports for PGS presents one set of verbal tags: uninformative (for LR = 1), limited support (2 to <100), moderate support (100 to <10,000), strong support (10,000 to <1,000,000), and very strong support ($\geq 1,000,000$).³⁸⁷ Analysts are allowed but not required to use these tags (and no others) along with the actual value.³⁸⁸ If any tag is used, then the examiner must list the full scale “to provide context for the numerical value.”³⁸⁹

384. Other analogies are possible. For example, an LR of about 1,000 is what you would get if someone flipped one of two coins—either a normal, evenly balanced coin or a trick coin with heads on both sides. You do not know which coin was flipped, but the evidence is that it came up heads every time in ten tosses. The evidence—the ten heads—is about 1,000 times more probable if it was the trick coin that was tossed than if it was the normal one. The expert would have to add that the probability that the trick coin was tossed also would depend on how likely it is that the person doing the flipping would choose the trick coin instead of the normal one—a factor that is beyond the data from the experiment on the coin (tossing it ten times). Similarly, the expert would have to say that she cannot give the probability that the defendant’s DNA was present. She can only report the probability *if* it was present compared to the probability *if* the alternative she considered was what had happened.

385. *E.g.*, United States v. Williams, 382 F. Supp. 3d 928, 934–35 (N.D. Cal. 2019) (“Based on SERI’s [Serological Research Institute’s] verbal scale, this result indicated ‘very strong’ support for the proposition that Elmore is a contributor to the sample.”) (footnote omitted).

386. *See, e.g.*, Raymond Marquis et al., *Discussion on How to Implement a Verbal Scale in a Forensic Laboratory: Benefits, Pitfalls and Suggestions to Avoid Misunderstandings*, 56 Sci. & Just. 364 (2016), <https://doi.org/10.1016/j.scijus.2016.05.009>.

387. U.S. Dep’t of Just., Uniform Language for Testimony and Reports for Forensic Autosomal DNA Examinations Using Probabilistic Genotyping Systems (Sept. 13, 2022) (link at <https://perma.cc/8VFG-KMHB>).

388. *Id.* at 3.

389. *Id.* The ULTR does not explain how dividing the continuum of possible LR values from zero to infinity into arbitrary segments provides meaningful context. As indicated in the section titled “Are Random-Match Probabilities That Are Smaller Than False-Positive Error Probabilities Irrelevant or Prejudicial?” of the third edition of this reference guide (Reference Manual on Scientific Evidence (3d ed. 2011)), some statisticians would argue that a more helpful explanation is to present an LR as the extent to which the evidence changes the odds in favor of a factual claim. *See also* Kaye & Stern, *supra* note 12, Appendix C.

DNA and RNA Typing of Bodily Fluids or Tissues

Standard DNA genotyping methods help answer the question of whose DNA is in the sample. In many cases, however, the more critical question is “how” rather than “who.” For example, a defendant accused of rape might concede that there was an assault but argue that the victim’s DNA found near the zipper of his pants resulted from her touching the fabric with her hands and was not from vaginal fluid, as the prosecution proposed.³⁹⁰ Forensic scientists call such factual claims activity-level propositions.³⁹¹ To supply activity-level evidence, molecular biology tests for differentiating between different types of bodily fluids and tissues are beginning to be used in casework.³⁹² Although the sequence of base pairs in DNA throughout the body is essentially identical, different parts of the coding DNA are active in each type of specialized cell that forms a tissue. Gene “expression” occurs when part of the double-stranded DNA molecule is transcribed into messenger RNA and that mRNA is translated into a protein.³⁹³ Some “housekeeping” proteins are in all tissues, but other proteins—and the associated mRNA transcripts—vary markedly in the extent to which they are present in different tissues. Skin cells have a different expression pattern than

390. *Cf.* *Holtzclaw v. State*, 448 P.3d 1134, 1148 (Okla. Ct. Crim. App. 2019) (no error for prosecutor to argue that DNA near the defendant’s zipper was “transferred by the vaginal fluids” even though the DNA analyst did not make that inference); Chong Wang et al., *Body Fluid Identification by Messenger RNA Profiling in Sexual Assault*, 8 J. Forensic Sci. & Med. 118 (2022), https://doi.org/10.4103/jfsm.jfsm_54_21 (“the suspect argued that the condom was just touched by the victim”).

391. R. Cook et al., *A Hierarchy of Propositions: Deciding Which Level to Address in Casework*, 38 Sci. & Just. 231 (1998), [https://doi.org/10.1016/S1355-0306\(98\)72117-3](https://doi.org/10.1016/S1355-0306(98)72117-3); Bas Kokshoorn et al., *Activity Level DNA Evidence Evaluation: On Propositions Addressing the Actor or the Activity*, 278 Forensic Sci. Int’l 115 (2017), <https://doi.org/10.1016/j.forsciint.2017.06.029>. To address the activity-level questions, forensic-science researchers and practitioners would like to bring to bear, in a standardized and systematic way, scientific studies of transfer, persistence, prevalence, and recovery of DNA. Duncan Taylor & Bas Kokshoorn, *Forensic DNA Trace Evidence Interpretation: Activity Level Propositions and Likelihood Ratios* (2023); Peter Gill et al., *DNA Comm’n of the Int’l Soc’y for Forensic Genetics: Assessing the Value of Forensic Biological Evidence—Guidelines Highlighting the Importance of Propositions. Part II: Evaluation of Biological Traces Considering Activity Level Propositions*, 44 Forensic Sci. Int’l: Genetics 102186 (2020), <https://doi.org/10.1016/j.fsigen.2019.102186>.

392. Andrea Patrizia Salzmann et al., *mRNA Profiling of Mock Casework Samples: Results of a FoRNAP Collaborative Exercise*, 50 Forensic Sci. Int’l: Genetics 102409 (2021), <https://doi.org/10.1016/j.fsigen.2020.102409>; Titia Sijen & SallyAnn Harbison, *On the Identification of Body Fluids and Tissues: A Crucial Link in the Investigation and Solution of Crime*, 12 Genes 1728, § 4.1 (2021), <https://doi.org/10.3390/genes12111728> (“RNA typing has been applied regularly at the Netherlands Forensic Institute since 2012”). This section discusses mRNA testing, but the related proteins themselves also can be analyzed with mass spectrometry. See, e.g., *Forensic Proteomics: Protein Identification and Profiling* (Eric D. Merkley ed., 2019).

393. See section titled “Genes and Gene Products” above.

liver cells, for example.³⁹⁴ Researchers have identified certain mRNA transcripts that are characteristic of forensically relevant fluids and thus can serve as markers for those fluids.

The mRNA in a biological sample can be extracted along with the DNA.³⁹⁵ Using the technologies for PCR amplification and capillary electrophoresis,³⁹⁶ or sequencing³⁹⁷ (see section above titled “What Are DNA Polymorphisms and How Are They Detected?”), the laboratory can try to determine the fluid or tissue in the trace-evidence sample. Other RNA transcripts than mRNA also have been proposed for fluid and tissue identification, as has testing for a chemical change to the DNA molecule that silences gene expression.³⁹⁸

Twins

Identical twins originate from a single fertilized egg cell (a zygote). Instead of multiplying to develop into a single embryo, fetus, and child, the developing set of cells breaks up into two or more masses that grow into separate children.³⁹⁹ Fraternal twins come from two different eggs fertilized by different sperm; those twins have many differences in their genomes. But in monozygotic twins, all the cells of the progeny come from the original zygote with a single genome. To the extent that the original chromosomes are faithfully copied in every cell

394. See Human Protein Atlas, The Tissue Section—Tissue-Based Map of the Human Proteome, <https://perma.cc/5UW9-UHNN>.

395. Amy D. Roeder & Cordula Haas, *Body Fluid Identification Using mRNA Profiling*, in Forensic DNA Typing Protocols 13 (William Goodwin ed., 2016), <https://doi.org/10.1007/978-1-4939-3597-0>.

396. Patricia Pearl Albani & Rachel Fleming, *Developmental Validation of an Enhanced mRNA-based Multiplex System for Body Fluid and Cell Type Identification*, 59 Sci. & Just. 217 (2019), <https://doi.org/10.1016/j.scijus.2019.01.001>; Tiffany R. Layne et al., *Rapid Microchip Electrophoretic Separation of Novel Transcriptomic Body Fluid Markers for Forensic Fluid Profiling*, 13 Micromachines 1657 (2022), <https://doi.org/10.3390/mi13101657>. To use a PCR-CE system, the mRNA first must be “reverse transcribed” into a DNA segment.

397. Cordula Haas et al., *Forensic Transcriptome Analysis Using Massively Parallel Sequencing*, 52 Forensic Sci. Int'l: Genetics 102486 (2021), <https://doi.org/10.1016/j.fsigen.2021.102486>; E. Hanson et al., *Messenger RNA Biomarker Signatures for Forensic Body Fluid Identification Revealed by Targeted RNA Sequencing*, 34 Forensic Sci. Int'l: Genetics 206 (2018), <https://doi.org/10.1016/j.fsigen.2018.02.020>.

398. Changes to DNA that do not alter the sequence of nucleotides are called “epigenetic.” The attachment of a small chemical group (a methyl group) on a gene stops it from producing proteins. The pattern of methylation of the DNA is different in different tissues. Athina Vidaki & Manfred Kayser, *Recent Progress, Methods and Perspectives in Forensic Epigenetics*, 37 Forensic Sci. Int'l: Genetics 180 (2018), <https://doi.org/10.1016/j.fsigen.2018.08.008>.

399. The monozygotic twinning rate is about 1 pregnancy per 250. Kurt Benirschke, *Multiple Gestation: The Biology of Twinning*, in Creasy and Resnik's Maternal-Fetal Medicine: Principles and Practice 53, 53 (Robert K. Creasy et al. eds., 7th ed. 2014).

division, monozygotic twins are genetically identical.⁴⁰⁰ Forensic STR analysis, which considers only tiny slices of the entire genome, cannot tell one from the other.

This limitation has thwarted prosecutions in which both twins were believed to have committed the crime⁴⁰¹ or in which an “evil twin” defense to an alleged crime committed by one twin was offered.⁴⁰² But variations in observable characteristics detectable by molecular tests related to DNA do exist.⁴⁰³ This section discusses one of them that has made an initial appearance in criminal and civil litigation.⁴⁰⁴ It has been proffered as “a realistic option, fit for practical forensic casework”⁴⁰⁵ that can support claims that one identical twin rather than the other is the source of trace-evidence DNA.⁴⁰⁶

Point mutations accumulate throughout life as cells split, giving rise to more and more distinguishable lines of cells. Even before birth, many “identical”

400. See *supra* note 25 & accompanying text.

401. In perhaps the most dramatic of these cases, German authorities did not bring charges in a “massive jewelry heist” because they could not decide which twin to charge. The German twins sent a message that they were “proud of the German constitutional state and gave it their thanks.” *Twins Suspected in Spectacular Jewelry Heist Set Free*, Spiegel Online Int’l, Mar. 19, 2009, <https://perma.cc/W6D8-GYH9>, last visited Mar. 9, 2023.

402. Alison Gee, *Twin DNA Test: Why Identical Criminals May No Longer Be Safe*, BBC News Mag., Jan. 15, 2014, <https://perma.cc/QM7N-DWWM>; Richard Willing, *Identical Twins Complicate Use of DNA Testing*, USA Today, Nov. 30, 2004, at A03.

403. There are differences in which genes are expressing proteins (“epigenetic” differences). Mario F. Fraga et al., *Epigenetic Differences Arise During the Lifetime of Monozygotic Twins*, 102 Proc. Nat’l Acad. Sci. 10604 (2005), <https://doi.org/10.1073/pnas.0500398102>. Forensic-science researchers have begun to study how to use this variation on samples of the same type of tissue from two twins. E.g., Benjamin Planterose Jiménez et al., *Equivalent DNA Methylation Variation Between Monozygotic Co-Twins and Unrelated Individuals Reveals Universal Epigenetic Inter-Individual Dissimilarity*, 22 Genome Biology 18 (2021), <https://doi.org/10.1186/s13059-020-02223-9> (“may one day allow universal epigenetic fingerprinting, which for instance is relevant in forensics for differentiating MZ twin individuals”); Athena Vidaki et al., *Epigenetic Discrimination of Identical Twins from Blood Under the Forensic Scenario*, 31 Forensic Sci. Int’l: Genetics 67 (2017), <https://doi.org/10.1016/j.fsigen.2017.07.014>. In addition, there are differences in the number of copies of a particular gene present in the genomes of different twins (“copy number variants”). Carl E.G. Bruder et al., *Phenotypically Concordant and Discordant Monozygotic Twins Display Different DNA Copy-Number-Variation Profiles*, 82 Am. J. Hum. Genetics 763 (2008), <https://doi.org/10.1016/j.ajhg.2007.12.011>. Furthermore, differences in mitochondrial sequences have been reported. Lijuan Yuan et al., *Identification of the Perpetrator Among Identical Twins Using Next Generation Sequencing Technology: A Case Report*, 44 Forensic Sci. Int’l: Genetics 102167 (2020), <https://doi.org/10.1016/j.fsigen.2019.102167>.

404. Burkhart Rolf & Michael Krawczak, *The Germlines of Male Monozygotic (MZ) Twins: Very Similar, But Not Identical*, 50 Forensic Sci. Int’l: Genetics 102408 (2021), <https://doi.org/10.1016/j.fsigen.2020.102408>.

405. Michael Krawczak et al., *Distinguishing Genetically Between the Germlines of Male Monozygotic Twins*, 14 PLoS Genetics e1007756 (2018), <https://doi.org/10.1371/journal.pgen.1007756>.

406. The technique has been used several times in Europe. *Id.*; Netherlands Forensic Institute, NFI Reports on Distinguishing Twins Using DNA Analysis, Oct. 19, 2022, <https://perma.cc/XWX8-83Y8>.

twins differ at some single base pairs (out of billions).⁴⁰⁷ After the embryo splits into twins, cells in each twin migrate and differentiate into distinct tissue types. If the mutation event occurs after twinning, it will be present only in one twin.⁴⁰⁸ If it occurs early, before the cells have specialized into all the cell types (nerve cells, blood cells, skin cells, germline cells, and so on), it may be perpetuated in multiple cell types and tissues. And if it occurs farther down the road, it will appear only in the tissue type of the first mutated cell.

To find distinguishing point mutations, a laboratory can perform whole-genome sequencing with an MPS method⁴⁰⁹ on a sample of blood from each twin. The twins are likely to have a small number of discordant base pairs.⁴¹⁰ Knowing the few discrepancies between the twins, the laboratory goes back to the trace-evidence DNA. Suppose the trace-evidence DNA came from a bloodstain. Sanger sequencing of that DNA at the distinguishing SNPs will show which twin's SNPs are also present in the trace-evidence DNA.

In this example, blood DNA has been compared to blood DNA. Suppose instead that the trace-evidence DNA is from semen, not blood, and the comparison DNA is from swabs of insides of the cheeks (the buccal mucosa) of the twins. The two types of collected cells developed from different parts of the embryo, and this fact complicates the analysis. Two or three weeks after fertilization, a set of cells in the embryo are slated to become germ cells (sperm or ova) in an event known as primordial germ cell specification (PGCS). Whether a mutation is propagated in sperm, in the mucosa, or in both—and whether these outcomes are present in one twin, the other twin, or both—depends on when (relative to PGCS and twinning) and where (within the embryo) they occurred.⁴¹¹ With cross-tissue comparisons, the technique thus requires establishing that at least some of the different SNPs in the twins' tissues are “twin-specific on the

407. Hakon Jonsson et al., *Differences Between Germline Genomes of Monozygotic Twins*, 53 *Nature Genetics* 27, 27 (2021), <https://doi.org/10.1038/s41588-020-00755-1> (“monozygotic twins differ on average by 5.2 early developmental mutations and . . . approximately 15% of monozygotic twins have a substantial number of these early developmental mutations specific to one of them”); Rui Li et al., *Somatic Point Mutations Occurring Early in Development: A Monozygotic Twin Study*, 51 *J. Med. Genetics* 28 (2014), <https://doi.org/10.1136/jmedgenet-2013-101712> (based on different SNPs found in 66 monozygotic twin pairs, “it is likely that each individual carries approximately over 300 postzygotic mutations in the nuclear genome of white blood cells that occurred in the early development”).

408. Conceivably, the same mutation could occur independently after twinning at the very same point within the billions of base pairs in a cell in the other twin. In that case, the single nucleotide change would not distinguish between the twins, but such a duplicated mutation is extremely unlikely.

409. See *supra* section titled “Sequencing and SNPs.”

410. To prune away sequencing errors with MPS, the laboratory can apply a slower but more accurate method such as Sanger sequencing to the small number of candidate variants.

411. See Jonsson et al., *supra* note 407.

one hand, but not (too) tissue-specific on the other.”⁴¹² In the semen-versus-cheek-swab situation, the premise is that the mutations occurred in early cells that gave rise to both progenitor germ cells and epithelial cells.

The evidentiary issues associated with this method of breaking the STR tie between twins include the validity of the sequencing methods;⁴¹³ the adequacy of the statistical model for calculating a pertinent likelihood ratio;⁴¹⁴ and the laboratory’s correct execution of the sequencing and statistical procedures. Also, a criminal-procedure issue would arise if one or both of the twins refused to give DNA samples. Compelling suspects to give bodily samples (breath, blood, saliva, hair, and fingerprint scrapings) is not unusual,⁴¹⁵ and warrants, nontestimonial orders, or subpoenas for DNA samples from third parties have been issued.⁴¹⁶ But these cases contemplated forensic STR or VNTR profiling, which reveals limited medical or other socially significant information.⁴¹⁷ Whole genome sequencing (WGS) reveals copious genetic information. Even if the only part of the genome that the government cares about and uses is a handful of SNPs that are useful only for differentiating between the two twins, it could be argued that compelling the twins to submit to the extraction of their blood or saliva for WGS is an unreasonable search or seizure.⁴¹⁸

412. Rolf & Krawczak, *supra* note 404. This wrinkle resulted in the exclusion of whole genome sequencing evidence in *Commonwealth v. McNair*, No. 8414CR10768 (Mass. Suffolk Cnty. Super. Ct., Apr. 11, 2017), *pet. for interlocutory rev. denied*, No. SJ-2017-0186 (Supreme Jud. Ct., Suffolk Cnty., July 27, 2017). The memorandum opinion is reprinted in David H. Kaye, *The Gordian Knot in Commonwealth v. McNair: What Did the Court Decide About Daubert?*, Forensic Sci., Stat. & L., Mar. 11, 2019, <https://perma.cc/3LJJ-P4K4>.

413. See *supra* section titled “Sequencing methods.”

414. Michael Krawczak et al., *Distinguishing Genetically Between the Germlines of Male Monozygotic Twins*, 14 PLoS Genetics e1007756 (2018), <https://doi.org/10.1371/journal.pgen.1007756>; David H. Kaye, *Identical Twins, Different Tissues, Disparate DNA*, Forensic Sci., Stat. & L., Mar. 7, 2019, <https://perma.cc/49AP-Q4WL>.

415. *Doe v. Senechal*, 725 N.E.2d 225 (Mass. 2000) (court-ordered buccal swab to test whether a member of the staff of a residential treatment facility for mentally ill adolescents fathered the child of a patient is a reasonable search and seizure); *In re Non-testimonial Identification Order Directed to R.H.*, 762 A.2d 1239 (Vt. 2000) (saliva sampling by court order based on reasonable suspicion).

416. *Bill v. Brewer*, 799 F.3d 1295 (9th Cir. 2015) (warrant for police officers’ DNA to exclude them as sources of crime-scene DNA); *In re G.B.*, 139 A.3d 885 (D.C. 2016) (warrant for a buccal swab from the victim of a stabbing for testing of DNA from blood droplets and stains at the apartment where the victim allegedly was stabbed). But cf. *Commonwealth v. Kostka*, 31 N.E.3d 1116 (Mass. 2015) (order to twin to provide a saliva sample to determine whether he was a fraternal as opposed to an identical twin was unreasonable where there was no showing of how definitive the profiling of a mixed sample from the victim’s fingernail was and no showing that defendant would raise an evil twin defense; and where fingerprint evidence could have excluded even an identical twin).

417. See *supra* note 39.

418. Cf. section titled “Investigative Genetic Genealogy” below. To forestall arguments over the validity or weight of cross-tissue comparisons, the prosecution in a rape case would need to

Investigative DNA Analysis

Much of this guide has presented DNA analysis as courtroom evidence that a criminal defendant is associated with a crime. The remaining sections describe DNA analysis from the perspective of police who have yet to zero in on a suspect. How can forensic scientists use DNA from a crime scene or a victim as a clue that could lead police to some actual suspects? Computerized databases of DNA profiles serve this function by allowing investigators to compare an unknown profile to millions of profiles from known individuals for “cold hits.” These law-enforcement intelligence databases are the subject of the next section. Police also can take advantage of large databases of other information on individual genomes maintained by companies to help members of the public locate possible relatives. The section below titled “Investigative Genetic Genealogy” describes the use of this genomic data to locate suspects indirectly, by first identifying possible relatives of the unknown source of the crime-scene DNA. A final section considers trace-evidence DNA, not purely as a token of individual identity or a step in constructing a family tree, but rather as an indication of biogeographic origin or external traits.

Offender and Suspect Database Searches

Law-enforcement databases and databanks

In 1989, Virginia became the first state to collect DNA from individuals convicted of certain crimes for a statewide DNA database—a set of records of DNA profiles that could be searched to see whether any of them matched a profile from biological evidence associated with a crime. Other states established similar systems,⁴¹⁹ and in 1998, the FBI launched software and hardware—the Combined DNA Index System (CODIS)—to allow nationwide searches to generate

secure semen samples from the twins. The balance of interests that would determine the constitutional reasonableness of compelling a twin to give such a sample would tilt further in the reluctant twin’s favor.

419. Although state and federal statutes authorizing and governing DNA databases for law enforcement are now pervasive, it is not obvious that the laws preclude a local police agency from maintaining its own records of DNA profiles from suspects they encounter or from biological evidence that they collect. “Rapid DNA” instruments and companies marketing software for database management have encouraged this practice. See *supra* section titled “Miniaturization for rapid capillary electrophoresis.” The “rogue databases” operating outside the constraints and quality assurance requirements of statutorily established systems have sparked concern. Murphy, *supra* note 287, at 180–88; N.Y. City Bar Ass’n Crim. Courts Comm., Crim. Just. Operations Comm., and Mass Incarceration Task Force, *Curbing Unregulated Local DNA Indexing* (Apr. 16, 2021).

an extremely valuable investigative lead: the name of an individual whose DNA, very likely, was left at the crime scene.⁴²⁰

CODIS has three layers. At the bottom, local governmental laboratories store the DNA profiles they obtain in a Local DNA Index System (LDIS). In the middle, a State DNA Index System (SDIS) holds these profiles and additional ones generated by statewide laboratories. State law specifies which profiles can be included. A National DNA Index System (NDIS) sits at the top. It houses SDIS profiles that meet federal requirements plus DNA profiles from federal laboratories. Approved law-enforcement laboratories can conduct database searches at any level—local, state, or national.⁴²¹

These CODIS databases have four main sets of records: a convicted offender index; an arrestee index; a detainee index;⁴²² and a forensic index (of profiles from DNA from crime scenes or left on victims).⁴²³ Matches can occur between either the convicted offender, arrestee, or detainee indexes and the forensic index. Cold hits also can occur within the forensic index, linking crimes to each other. The underlying samples (or just the DNA extracts) also are retained, making the system a databank as well as an electronic database.⁴²⁴ Finally, the system as a whole includes “population databases.” These are anonymized research databases used to estimate allele frequencies that go into the equations

420. A pilot project had been started in 1990. U.S. Dep’t of Just., Office of the Inspector General, Combined DNA Index System Operational and Laboratory Vulnerabilities, Audit Report 06–32 (2006). The DNA Identification Act of 1994, 34 U.S.C. §§ 40702–40703 (later amended in various respects), authorized the FBI to create a national database. At the outset of the fully operational system, only nine states participated, and the federal government did not collect DNA from federal offenders. Nathan James, Cong. Reference Serv., DNA Testing in Criminal Justice: Background, Current Law, Grants, and Issues 3 (2014) (CRS Rep. R41800).

421. On the requirements for uploading a crime-scene profile to NDIS so that it will be searched against the offender and arrestee databases, see *Cowels v. FBI*, 936 F.3d 62 (1st Cir. 2019) (FBI’s refusal to accept a sample that a state court ordered uploaded to NDIS at the request of defendants, who had been convicted but were granted a new trial, was not arbitrary and capricious under the Administrative Procedure Act where it was not clear that the sample was connected to the crime). The scope of the information maintained at each level of the hierarchy of databases varies. LDIS has names and crime information that are not uploaded to SDIS and NDIS.

422. Detainees are “non-United States persons who are detained under the authority of the United States.” 28 C.F.R. § 28.12(b) (2023). “Non-United States persons” means “persons who are not United States citizens and who are not lawfully admitted for permanent residence.” *Id.* DNA is not routinely collected from “(1) [a]liens lawfully in, or being processed for lawful admission to, the United States; (2) [a]liens held at a port of entry during consideration of admissibility and not subject to further detention or proceedings; or (3) [a]liens held in connection with maritime interdiction.” *Id.* On the rationale for these exceptions, see 85 Fed. Reg. 13483 (Mar. 9, 2020).

423. It also has profiles from unidentified remains and from samples voluntarily provided by relatives of missing persons. Butler, *supra* note 43, at 236.

424. *Id.* at 214–15; *United States v. Kriesel*, 720 F.3d 1137 (9th Cir. 2013) (describing the reasons for sample retention and holding that Fed. R. Crim. 41(g) does not require the government to return the blood sample after an offender has completed his sentence and term of supervised release).

for genotype frequencies, random-match probabilities, or likelihood ratios.⁴²⁵ They come from other sources and are not searched for cold hits.⁴²⁶

Law-enforcement DNA databases and databanks have provoked considerable litigation. Fourth Amendment challenges to compulsory DNA collection from convicted offenders almost always failed,⁴²⁷ although opinions were divided as to the appropriate doctrinal route around normal warrant and probable cause requirements,⁴²⁸ and several vigorous dissenting opinions emerged.⁴²⁹ Expanding the scope of the databases to include profiles from individuals who were merely indicted or arrested produced a split of authority in the lower courts. Eventually, the Supreme Court, in *Maryland v. King*,⁴³⁰ narrowly upheld a state statute requiring all custodial arrestees to submit to DNA sampling as part of the booking process for certain violent crimes and burglary.⁴³¹

425. These quantities are discussed *supra* in the sections titled “Inference, Statistics, and Population Genetics in Human Nuclear DNA Testing” and “(1) Definition of the likelihood ratio”.

426. Butler, *supra* note 43, at 215. The failure to recognize the difference between searchable records for possible cold hits and de-identified databases for population frequency research led the state of Texas to destroy its public health system’s repository of newborn blood samples. David H. Kaye, *Up in Smoke: 5 Million Neonatal Blood Samples Incinerated*, Forensic Sci., Stat. & L., July 27, 2014, <https://perma.cc/2LN2-NQXL>. On law-enforcement access to newborn samples collected for genetic-disease screening, see Natalie Ram, *America’s Hidden National DNA Database*, 100 Tex. L. Rev. 1253 (2022).

427. See *United States v. Kincade*, 379 F.3d 813 (9th Cir. 2004) (en banc); Robin Cheryl Miller, *Validity, Construction, and Operation of State DNA Database Statutes*, 76 A.L.R. 5th 239 (2000).

428. See *Banks v. United States*, 490 F.3d 1178, 1183 (10th Cir. 2007) (“our sister circuits have taken different analytical routes to analyzing DNA-indexing statutes” and “our own precedents are divided”); Padgett v. Donald, 401 F.3d 1273, 1278 (11th Cir. 2005) (“Each circuit to address the question has upheld the constitutionality of DNA profiling statutes, but the circuits have disagreed on whether to do so through the special needs analysis or through the traditional balancing test.”).

429. *Kincade*, 379 F.3d at 842 & 871 (dissenting opinions).

430. 569 U.S. 435 (2013).

431. *King* has been extended to uphold other database laws that encompass more crimes, that do not automatically remove profiles and samples if no conviction follows the arrest, and that can be used for “familial searching.” *Haskell v. Harris*, 745 F.3d 1269 (9th Cir. 2014) (en banc) (per curiam); *Haskell v. Brown*, 317 F. Supp. 3d 1095 (N.D. Cal. 2018); *People v. Buza*, 413 P.3d 1132 (Cal. 2018) (but defendant’s alleged crime fell within the “serious” category of *King*); cf. *State v. Medina*, 102 A.3d 661 (Vt. 2014) (striking down post-arraignment, pre-conviction, DNA sampling under the Vermont constitution). For academic debate on the reach of *King* and the Fourth Amendment logic of the majority and the dissent, compare Erin Murphy, *License, Registration, Cheek Swab: DNA Testing and the Divided Court*, 127 Harv. L. Rev. 161 (2013), with David H. Kaye, *Maryland v. King: Per Se Unreasonableness, the Golden Rule, and the Future of DNA Databases*, 127 Harv. L. Rev. Forum 39 (2013). See also David H. Kaye, *Why So Contrived? DNA Databases After Maryland v. King*, 104 J. Crim. L. & Criminology 535 (2014); Tracey Maclin, *Maryland v. King: Terry v. Ohio Redux*, 2013 Sup. Ct. Rev. 359; Andrea Roth, *Maryland v. King and the Wonderful, Horrible DNA Revolution in Law Enforcement*, 11 Ohio St. J. Crim. L. 295 (2013).

Probative value of database hits

If the DNA profile from a trace-evidence sample matches a profile in an arrestee, detainee, or offender database, the person identified by the cold hit will become the target of the investigation. Prosecution may follow. These database-trawl cases can be contrasted with traditional “confirmation cases” in which the defendant already was a suspect and the DNA testing provided additional evidence. In confirmation cases, statistics such as the estimated frequency of the matching DNA profile in various populations, the equivalent random-match probabilities, or likelihood ratios can be used to indicate the probative value of the DNA match.⁴³²

In trawl cases, however, an additional question arises: Does the fact that the defendant was selected for prosecution by trawling require some adjustment to the usual statistics? The legal issues are twofold. First, is a particular quantity—be it the unadjusted random-match probability or some adjusted probability—scientifically valid (or generally accepted) in the case of a database search? If not, it must be excluded under the *Daubert* (or *Frye*) standards. Second, is the statistic irrelevant or unduly misleading? If so, it must be excluded under the rules that require all evidence to be relevant and not unfairly prejudicial. To clarify, we summarize the statistical literature on this point. Then, we describe the case law.

The statistical analyses of adjustment. All statisticians agree that, in principle, the search strategy affects the probative value of a DNA match. One group describes and emphasizes the impact of the database match on the hypothesis that the database does not contain the source of the crime-scene DNA. This is a “frequentist” view. It asks how frequently searches of innocent databases—those for which the true source is someone outside the database—will generate cold hits. From this perspective, trawling is a form of “data mining” that produces a “selection effect” or “ascertainment bias.”⁴³³ If we pick a lottery ticket at random, the probability p that we have the winning ticket is negligible. But if we search through all the tickets, sooner or later we will find the winning one. And even if we search through some smaller number N of tickets, the probability of picking a winning ticket is no longer p , but Np .⁴³⁴

432. On the computation and admissibility of such statistics, see *supra* sections titled “Inference, Statistics, and Population Genetics in Human Nuclear DNA Testing” and “Likelihood ratio (LR)” (in the section titled “Mixtures”).

433. See Kaye & Stern, *supra* note 12.

434. The analysis of the DNA database search is more complicated than the lottery example suggests. In the simple lottery, there was exactly one winner. The trawl case is closer to a lottery in which we hold a ticket with a winning number, but it might be counterfeit, and we are not sure how many counterfeit copies of the winning ticket were in circulation when we bought our N tickets.

Likewise, if DNA from N innocent people is examined to determine if any of them match the crime-scene DNA, then the probability of a match in this group is not p , but some quantity that could be as large as Np . This type of reasoning led the 1996 NRC committee to recommend that “[w]hen the suspect is found by a search of DNA databases, the random-match probability should be multiplied by N , the number of persons in the database.”⁴³⁵ The 1992 committee⁴³⁶ and the FBI’s former DNA Advisory Board⁴³⁷ took a similar position.

No one questions the mathematics that show that when the database size N is very small compared with the size of the population, Np is an upper bound on the expected frequency with which searches of databases will incriminate innocent individuals when the true source of the crime-scene DNA is not represented in the databases. The Bayesian school of thought, however, suggests that the frequency with which innocent databases will be falsely accused of harboring the source of the crime-scene DNA is basically irrelevant. The question of interest to the legal system is whether the one individual whose database DNA matches the trace-evidence DNA is the source of that trace. As the size of a database approaches that of the entire population, finding one and only one matching individual should be more, not less, convincing evidence against that person. Thus, instead of looking at how surprising it would be to find a match in a large group of innocent suspects, this school of thought asks how much the result of the database search enhances the probability that the individual so identified is the source. The database search is actually more probative than the confirmation search because the DNA evidence in the trawl case is much more extensive. Trawling through large databases excludes millions of people, thereby reducing the number of people who might have left the trace evidence if the suspect did not. This additional information increases the likelihood that the defendant is the source, although the effect is indirect and generally small.⁴³⁸

435. NRC II, *supra* note 7, at 161 (Recommendation 5.1).

436. The first NRC committee wrote that “[t]he distinction between finding a match between an evidence sample and a suspect sample and finding a match between an evidence sample and one of many entries in a DNA profile databank is important.” It used the same Np formula in a numerical example to show that “[t]he chance of finding a match in the second case is considerably higher, because one . . . fishes through the databank, trying out many hypotheses.” NRC I, *supra* note 6, at 124. Rather than proposing a statistical adjustment to the match probability, however, that committee recommended using only a few loci in the databank search, then confirming the match with additional loci, and presenting only “the statistical frequency associated with the additional loci.” *Id.* at 124 tbl. 1.1.

437. DNA Advisory Bd., Statistical and Population Genetics Issues Affecting the Evaluation of the Frequency of Occurrence of DNA Profiles Calculated from Pertinent Population Database(s) (July 2000), <https://perma.cc/JQ4C-6ZGR>.

438. When the size of the database approaches the size of the entire population, the effect is large. Also, finding that only one individual in a large database has a particular profile raises the probability that this profile is very rare, further enhancing the probative value of the DNA evidence.

Of course, when the cold hit is the only evidence against the defendant, the total package of evidence in the trawl case is less than in the confirmation case. Nonetheless, the Bayesian treatment shows that the DNA part of the total evidence is more powerful in a cold-hit case because this part of the evidence is more complete than when the search for matching DNA is limited to a single suspect. This reasoning suggests that the random-match probability (or, equivalently, the frequency p in the population) understates the probative value of the unique DNA match in the trawl case. And if this is so, then Np is misleadingly large, and the unadjusted random-match probability or frequency p can be used as a conservative indication of the probative value of the finding that, of the many people in the database, only the defendant matches.⁴³⁹

Judicial opinions on adjustment. The need for an adjustment has been vigorously debated in the statistical literature, and to a lesser extent in the legal literature.⁴⁴⁰ The dominant view in the journal articles is that the random-match probability, frequency, or likelihood ratio need not be inflated to protect the defendant—at least, not as a matter of mathematics.⁴⁴¹ The major opinions to

439. This analysis was developed by David Balding and Peter Donnelly. For informal expositions, see, for example, Peter Donnelly & Richard D. Friedman, *DNA Database Searches and the Legal Consumption of Scientific Evidence*, 97 Mich. L. Rev. 931 (1999); David H. Kaye, *Rounding Up the Usual Suspects: A Legal and Logical Analysis of DNA Database Trawls*, 87 N. Car. L. Rev. 425 (2009); Simon Walsh et al., *DNA Intelligence Databases*, in *Forensic DNA Evidence Interpretation* 429, 443–44 (John Buckleton et al. eds., 2d ed. 2016); Simon Walsh & John Buckleton, *DNA Intelligence Databases*, in *Forensic DNA Evidence Interpretation* 439, 465–68 (John Buckleton et al. eds., 2005). For a related analysis directed at the average probability that an individual identified through a database trawl is the source of a crime-scene DNA sample, see Yun S. Song et al., *Average Probability That a “Cold Hit” in a DNA Database Search Results in an Erroneous Attribution*, 54 J. Forensic Sci. 22, 23–24 (2009), <https://doi.org/10.1111/j.1556-4029.2008.00917.x>. Yet another variation is in Ian Ayres & Barry Nalebuff, *The Rule of Probabilities: A Practical Approach for Applying Bayes’ Rule to the Analysis of DNA Evidence*, 67 Stan. L. Rev. 1447 (2015); cf. John T. Wixted et al., *Calculating the Posterior Odds from a Single-match DNA Database Search*, 18 L. Probability & Risk 1 (2018), <https://doi.org/10.1093/lpr/mgz001> (with commentary).

440. For citations to this literature, see Kaye, *supra* note 439; Ronald W.J. Meester & Klaas Slooten, *DNA Database Matches: A p Versus np Problem*, 46 Forensic Sci. Int’l: Genetics 102229 (2020); <https://doi.org/10.1016/j.fsigen.2019.102229>; Walsh et al., *supra* note 439, at 443.

441. Marjan J. Sjerps, *Probabilistic Considerations when Interpreting Database Searches and Selection Effects*, in *Handbook of Forensic Statistics* 325, 331 (David Banks et al. eds., 2021) (“[N]umerous papers have been published after Donnelly and Friedman [*supra* note 439] and [A. Philip Dawid, *Comment on Stockmarr’s “Likelihood Ratios for Evaluating DNA Evidence When the Suspect Is Found Through a Database Search,”* 57 Biometrics 976 (2001)], exploring the controversy with, e.g., Bayesian networks, or rewriting formulas in a more convenient way, reaching for the most part the same conclusion. . . . Thus, from a mathematical point of view, the database search controversy has been solved.”). On the manner and situations in which p , Np , or some combination or variation of the two should be presented, there is less agreement. See ENFSI DNA Working Group, *DNA Database*

confront this issue agree that p is admissible against the defendant in a trawl case.⁴⁴² They reason that all the statistics are admissible under *Frye* and *Daubert* because there is no controversy over how they are computed. They then assume that both p and Np are logically relevant and not prejudicial in a trawl case.

Kinship trawling (“familial searching”)

Normally, police trawl a DNA database to see if any recorded STR profiles match a crime-scene profile. On occasion, these trawls could lead to charges against a defendant whose DNA is not even in the databank. This can occur for identical twins, one of whom is a convicted offender whose DNA type is on file and the other who has never been typed. If the convicted twin was in prison when the crime under investigation was committed, and if the police realize that he had an identical twin, suspicion should fall on the twin. Presumably, the police would seek a sample from this twin, and at trial it would not be necessary for the prosecution to explain the roundabout process through which the defendant was identified.

In this hypothetical example, the defendant fortuitously was found because a relative’s DNA led the police to him. More generally, the fact that close relatives share more alleles than other members of the same subpopulation can be exploited as a deliberate investigative tool. Rather than search for a match at all the loci in an STR profile, police could search for a near miss—a partial match that is much more probable when the partly matching profile in the database comes from a close relative than when it comes from an unrelated person.⁴⁴³

Management Review and Recommendations (Apr. 2019), accessible via European Network of Forensic Science Institutes (ENFSI) Best Practice Manuals and Forensic Guidelines, <https://perma.cc/XN6B-8DAY>; David H. Kaye, *People v. Nelson: A Tale of Two Statistics*, 7 *Law, Probability & Risk* 249 (2008), <https://doi.org/10.1093/lpr/mgn005>; Kaye, *supra* note 439; Meester & Slooten, *supra* note 440; Murphy, *supra* note 287, at 118; Sjerps, *supra*; Wixted et al. (and commentary), *supra* note 439.

442. *People v. Nelson*, 185 P.3d 49 (Cal. 2008); *United States v. Jenkins*, 887 A.2d 1013 (D.C. 2005). The cases are analyzed in Kaye, *supra* note 439 & note 441. More recent cases are similar. *E.g.*, *People v. Turner*, 476 P.3d 676 (Cal. 2020); *Commonwealth v. Bizanowicz*, 945 N.E.2d 356, 362–63 (Mass. 2011). In *Turner*, the California Supreme Court took a step toward the pure random-match probability approach. 476 P.3d at 693 & 693 n.12 (intimating that the Np statistic would not be “germane” when p is extremely small, but it could be admitted “in some circumstances” when “a profile’s random match probability [is] relatively high”). Why or how to draw this line is not clear.

443. Frederick R. Bieber et al., *Finding Criminals Through DNA of Their Relatives*, 312 *Science* 1315 (2006), <https://doi.org/10.1126/science.1122655>. The FBI distinguishes between kinship trawling to locate an individual outside the database and “partial match” searching to locate an individual inside the database when the crime-scene profile is incomplete. FBI, *supra* note 9 (questions 25 & 27–32). The latter can incidentally produce a lead to someone outside the database, but it is not an efficient procedure for that purpose.

Analysis of Y-STRs or mtDNA from the DNA sample in the databank then could determine whether the “database inhabitant” probably is in the same paternal or maternal lineage as the unknown individual who left DNA at the scene of the crime. Apparent relatives who emerge from this process become candidates for further investigation based on nongenetic information such as age and location. (Alternatively, a Y-STR or mtDNA database could be queried at the outset for possible matches to the trace-evidence DNA. Several countries have used this strategy.⁴⁴⁴)

Such “familial searching”⁴⁴⁵ raises technical and policy questions. The technical and practical challenge is to devise a search strategy that detects the relatives who happen to be in the database while keeping the number of false leads to a tolerable level. For autosomal STR profiling supplemented by Y-STR or mtDNA filtering, software that computes a “kinship index” for STR profiles skims the database for first-degree relatives (parent–child and sibling relationships) based on the pattern of allele sharing and the frequency of the shared alleles.⁴⁴⁶

The policy questions are whether exposing relatives to the possibility of being investigated on the basis of genetic leads from their kin is fair and whether the benefits of kinship trawling outweigh the dangers of intrusions from false

444. Jianye Ge & Bruce Budowle, *Forensic Investigation Approaches of Searching Relatives in DNA Databases*, 66 J. Forensic Sci. 430 (2021), <https://doi.org/10.1111/1556-4029.14615> (explaining the advantages and noting the costs of this form of kinship trawling).

445. On this term and possibly clearer alternatives to it, see David H. Kaye, *The Genealogy Detectives: A Constitutional Analysis of “Familial Searching,”* 50 Am. Crim. L. Rev. 109, 120–21 (2013). “Familial searching” became prominent in the United States with the conviction of the serial killer dubbed the Grim Sleeper, who seemed to take long breaks between murders of young women in South Los Angeles from the mid-1980s until at least 2007. Mike McPhate, *Los Angeles Man Is Convicted in ‘Grim Sleeper’ Serial Killer Case*, N.Y. Times, May 5, 2016. California authorities located him via the DNA profile of his son, whose profile was in the state database as the result of a 2008 arrest for firearm and drug offenses. Marisa Gerber, *The Controversial DNA Search that Helped Nab the ‘Grim Sleeper’ Is Winning over Skeptics*, L.A. Times, Oct. 25, 2015. A subsequent and even more highly publicized development—involving trawls of nongovernmental databases for large blocks of SNPs that reveal much more distant relationships—is discussed *infra* in the section titled “Investigative Genetic Genealogy.” For clarity, the term “familial searching” should be reserved for the trawls, like the one in the Grim Sleeper case, for close relatives in law-enforcement databases.

446. The database profiles with the largest kinship-index values rise to the top. A siblingship index, a paternity index, or any other kinship index is a kind of likelihood ratio (a term explained *supra* section titled “(1) Definition of the likelihood ratio”). It contrasts the probability of observing the STR types when the profile in the database and the profile of the unknown source of the crime-scene sample have a certain relationship (for example, siblingship) to the probability of observing these types when they are unrelated. Jianye Ge et al., *Comparisons of Familial DNA Database Searching Strategies*, 56 J. Forensic Sci. 1448, 1448 (2011), <https://doi.org/10.1111/j.1556-4029.2011.01867.x>; Steven P. Myers et al., *Searching for First-Degree Familial Relationships in California’s Offender DNA Database: Validation of a Likelihood Ratio-Based Approach*, 5 Forensic Sci. Int’l: Genetics 493 (2011), <https://doi.org/10.1016/j.fsigen.2010.10.010>.

leads, disruption of family harmony, and exposure of family or individual secrets, such as misattributed paternity.⁴⁴⁷ The extent to which such concerns could support constitutional challenges to kinship trawling has been debated.⁴⁴⁸ A few states have banned kinship trawling of their law-enforcement databases, and several jurisdictions subject it to “an approval process, specified and limited offense types, and an affirmative assertion by specific stakeholders that all leads have been exhausted.”⁴⁴⁹

All-pairs matching within a database to verify estimated random-match probabilities

A third and final use of police intelligence databases has evidentiary implications. The section above titled “Frequencies, Probabilities, and Prejudice” explained the use of population-genetics models to estimate DNA-genotype frequencies. Large databases can be used to check these theoretical computations. Using the New Zealand national database, which consisted of six-locus profiles, for example, researchers compared every profile with every other profile.⁴⁵⁰ At the time, there were 10,907 profiles in the database. Although this number is not huge, 10,907 profiles can be arranged to form about 59 million distinct pairs.⁴⁵¹ Because the theoretical random-match probability was about 1 in 50 million, if all the individuals represented in the database were unrelated, one would expect that an exhaustive comparison of the profiles for these

447. See, e.g., Henry T. Greely et al., *Family Ties: The Use of DNA Offender Databases to Catch Offenders’ Kin*, 34 J.L. Med. & Ethics 248 (2006), <https://doi.org/10.1111/j.1748-720X.2006.00031.x>; Erica Haimes, *Social and Ethical Issues in the Use of Familiar Searching in Forensic Investigations: Insights from Family and Kinship Studies*, 34 J.L. Med. & Ethics 263 (2006), <https://doi.org/10.1111/j.1748-720X.2006.00032.x>; Erin Murphy, *Relative Doubt: Familial Searches of DNA Databases*, 109 Mich. L. Rev. 291 (2010); Murphy, *supra* note 287, at 189–214; Sonia M. Suter, *All in the Family: Privacy and DNA Familial Searching*, 23 Harv. J.L. & Tech. 309 (2010).

448. Compare Murphy, *supra* note 447, and Murphy, *supra* note 287, at 208–11, with Kaye, *supra* 445.

449. Ge & Budowle, *supra* note 444, at 432; see also Tepring Piquado et al., *Forensic Familial and Moderate Stringency DNA Searches: Policies and Practices in the United States, England, and Wales* (2019), <https://perma.cc/8HPE-V32Y>.

450. Walsh & Buckleton, *supra* note 439, at 463.

451. For the people represented in the New Zealand database, about $10,907 \times 10,907$ pairs—such as (1,1), (1,2), (1,3), . . . , (1,10907), (2,1), (2,2), (2,3), . . . , (2,10907), . . . , (10907,1), (10907,2), (10907,3), (10907,10907)—can be formed. This amounts to almost 119 million possible pairs. But there is no point in checking the pairs (1,1), (2,2), . . . (10,907,10,907)—a profile obviously matches itself. Thus, the number of ordered pairs with different individuals is 119 million minus 10,907. Finally, ordered pairs such as (1,5) and (5,1) involve the same two people. Therefore, the number of *distinct* pairs of people is about half of 119 million—the 59 million figure in the text. More generally, an all-pairs search in a large database of size N involves $N(N-1)/2$, or about $N^2/2$ comparisons.

59 million pairs would produce only about one match. In fact, the 59 million comparisons revealed 10 matches. What could have caused this excess number of matches? Further investigation showed that eight of the pairs were twins or brothers, the ninth was a duplicate (because one person gave a sample as himself and then again pretending to be someone else), and the tenth was apparently a match between two unrelated people. This exercise thus confirmed the theoretical computation of the random-match probability. On average, the theoretical match probability was about 1/50,000,000, and the rate of matches in the unrelated pairs within the database was 1/59,000,000.

In the United States, defendants have argued that the unintuitively high number of partial matches from internal all-pairs trawls of offender databases demonstrates that the usual estimates of random-match probabilities are incorrect,⁴⁵² and they have sought discovery of the criminal-offender databases to determine whether the number of matching and partially matching pairs exceeds the predictions made with the population-genetics model.⁴⁵³ An early report about partial matches (at any nine loci out of the 13 CODIS core loci) in the state database in Arizona was said to show extraordinarily large numbers of partial matches (without accounting for the combinatorial explosion in the number of comparisons in an all-pairs database search).⁴⁵⁴ However, some

452. *E.g.*, *People v. Luna*, 989 N.E.2d 655, 664–65 (Ill. App. Ct. 2013) (defense expert's testimony on results of an all-pairs trawl of the Illinois state dataset).

453. *United States v. Davis*, 602 F. Supp. 2d 658, 681 (D. Md. 2009) (reporting results from an all-pairs search of the Maryland database ordered by a state court); *Derr v. State*, 73 A.3d 254 (Md. 2013) (defendant who matched at 13 loci was not entitled to have the FBI determine how many pairwise matches there would be at 9 through 13 loci; neither was he entitled to all the profiles in the national database so that his experts could conduct these pairwise comparisons); *State v. Dwyer*, 985 A.2d 469 (Me. 2009) (defendant was not entitled to require the state to perform an all-pairs search of its database for matches at nine or more loci); *Young v. United States*, 63 A.3d 1033 (D.C. 2013) (defendant who matched at 13 loci was not entitled to all-pairs searching of the national database); Jason Felch & Maura Dolan, *How Reliable Is DNA in Identifying Suspects?*, L.A. Times, July 20, 2008.

454. Felch & Dolan, *supra* note 453; David H. Kaye, *Trawling DNA Databases for Partial Matches: What Is the FBI Afraid Of?*, 19 Cornell J.L. & Pub. Pol'y 145 (2009); Murphy, *supra* note 287, at 107–08. As illustrated, *supra* note 451, an all-pairs search in a large database of size N involves nearly $N^2/2$ comparisons. For example, a database of 6 million samples gives rise to some 18,000,000,000,000 comparisons. Even with no population structure, relatives, and duplicates, and with random-match probabilities in the trillionths, one would expect to find a large number of matches or near-matches. An analogy can be made to the famous “birthday problem.” In its simplest form, the birthday problem assumes that equal numbers of people are born every day of the year. The problem is to determine the minimum number of people in a room such that the odds favor there being at least two of them who were born on the same day of the same month. Focusing solely on the random-match probability of 1/365 for a specified birthday makes it appear that a huge number of people must be in the room for a match to be likely. After all, the chance of a match between two individuals having a given birthday (say, January 1) is (ignoring leap years) a minuscule $1/365 \times 1/365 = 1/133,225$. But because the matching birthday can be any

scientists question the utility of the investigative databases for population-genetics research.⁴⁵⁵ They observe that these databases contain an unknown number of relatives, that they might contain duplicates, and that the population in the offender databases is highly structured. These complicating factors would need to be considered in testing for an excess of matches or partial matches. Studies of offender databases in Australia and New Zealand that made adjustments for population structure and close relatives showed substantial agreement between the expected and observed numbers of partial matches, at least up to the nine STR loci used for database profiles in those countries at the time of the studies.⁴⁵⁶

The existence of large databases also provides a means of estimating a random-match probability without making any modeling assumptions. For the New Zealand study, even ignoring the possibility of relatives and duplicates, there were only 10 matches out of 59 million comparisons. The empirical estimate of the random-match probability is therefore about 1 in 5.9 million. This is about 10 times larger than the theoretical estimate, but still quite small. As this example indicates, crude but simple empirical estimates from all-pairs comparisons in large databases may well produce random-match probabilities that are larger than the theoretical estimates (as expected when full siblings or other close relatives are in the databases), but the estimated probabilities are likely to remain impressively small.

Investigative Genetic Genealogy (IGG)

Until recently, personal knowledge, oral histories, and documents were the only ways to connect extended family members within and across generations. Since

one of the 365 days in the year and because there are $N(N-1)/2$ ways to have a match, it takes only $N = 23$ people before it is more likely than not that at least two people share a birthday. The birthday problem thus shows that surprising coincidences commonly occur even in relatively small databases. See, e.g., Persi Diaconis & Frederick Mosteller, *Methods for Studying Coincidences*, 84 J. Am. Stat. Ass'n 853 (1989).

455. Bruce Budowle et al., *Partial Matches in Heterogeneous Offender Databases Do Not Call into Question the Validity of Random Match Probability Calculations*, 123 Int'l J. Legal Med. 59 (2009), <https://doi.org/10.1007/s00414-008-0239-1>.

456. James M. Curran et al., *Empirical Support for the Reliability of DNA Evidence Interpretation in Australia and New Zealand*, 40 Australian J. Forensic Sci. 99, 102–06 (2008), <https://doi.org/10.1080/00450610802172230>; Bruce S. Weir, *The Rarity of DNA Profiles*, 1 Annals Applied Stat. 358 (2007), <https://doi.org/10.1214/07-AOAS128>; Bruce S. Weir, *Matching and Partially-Matching DNA Profiles*, 49 J. Forensic Sci. 1009, 1013 (2004); cf. Laurence D. Mueller, *Can Simple Population Genetic Models Reconcile Partial Match Frequencies Observed in Large Forensic Databases?*, 87 J. Genetics (India) 101 (2008), <https://doi.org/10.1007/s12041-008-0016-4> (maintaining that excess partial matches in an Arizona offender database are not easily reconciled with theoretical expectations). This literature is reviewed in Kaye, *supra* note 454.

2007, direct-to-consumer genetic testing (DTC-GT) has made it possible to find relatives through inherited DNA. Tens of millions of people have undergone such testing for information about their ancestral origin and potential biological relation to other people within the vendors' databases.⁴⁵⁷ Law-enforcement officials also have turned to some of these databases to track down the source of crime-scene DNA or to identify human remains indirectly. Like the members of the public interested in finding possible relatives, they try to learn from the companies that assemble these databases if DNA data from potential relatives of the source of the forensic sample seem to be included in the genealogy databases. Constructing parts of the family tree of these relatives has led investigators to the source of the forensic DNA in hundreds of cases.⁴⁵⁸ This section explains how this adaptation of personal genetic genealogy is conducted, how it differs from familial searching with law-enforcement databases, how it might be regulated, and how it might implicate the Fourth Amendment. We begin with a brief introduction to the use of DTC-GT databases in genealogy research by members of the public.

How does direct-to-consumer genetic genealogy work?

A DTC-GT company supplies paying customers with mail-in spit kits or cheek swabs. Upon receiving the sample, the company uses certain microarrays (the SNP chips described in the section above titled "Sequence-Specific Probes and Microarrays") to type hundreds of thousands of SNPs. These SNPs are spaced across the 22 autosomal chromosomes, somewhat like individual letters sampled from the volumes of an encyclopedia without the intervening text. Thus, they are not themselves DNA sequences.

Starting in 2009, DTC-GT companies began assembling this information into databases to assist customers seeking to discover possible relatives.⁴⁵⁹ A DTC-GT customer can have the company trawl the database with specialized software to identify the chromosomal locations and lengths of blocks of SNPs (haploblocks) that match between the customer's data file and the files

457. James W. Hazel et al., *Direct-to-Consumer Genetic Testing: Prospective Users' Attitudes Toward Information About Ancestry and Biological Relationships*, 16 PLoS One e0260340 (2021), <https://doi.org/10.1371/journal.pone.0260340>. "This genetic genealogy has enabled thousands of individuals who have lost their biological identity through adoption, abandonment, anonymous gamete donation, misattributed parentage, etc., to regain their genetic heritage." Ellen M. Greytak et al., *Genetic Genealogy for Cold Case and Active Investigations*, 299 Forensic Sci. Int'l 103, 103 (2019), <https://doi.org/10.1016/j.forsciint.2019.03.039>.

458. Tracey L. Dowdeswell, *Forensic Genetic Genealogy: A Profile of Cases Solved*, 58 Forensic Sci. Int'l: Genetics 102679 (2022), <https://doi.org/10.1016/j.fsigen.2022.102679>.

459. Daniel Kling et al., *Investigative Genetic Genealogy: Current Methods, Knowledge and Practice*, 52 Forensic Sci. Int'l: Genetics 102474, at 2 (2021), <https://doi.org/10.1016/j.fsigen.2021.102474>.

in the company's database.⁴⁶⁰ Large matching haploblocks—measured in centimorgans⁴⁶¹—are believed to be “identical by descent” (IBD) and thus indicative of a common recent ancestor.⁴⁶² As a rule, the greater the total length of the shared IBD haploblocks, the closer the relationship.⁴⁶³ The total length thus indicates “the probability that the relationship between the unknown individual and the match falls into each degree of relatedness.”⁴⁶⁴ But different relationships can correspond to the same range of haploblock sharing.⁴⁶⁵ For example, a total of 100 centimorgans could indicate “anywhere from 5th degree to >8th degree, with 6th degree being most likely.”⁴⁶⁶ Consequently, the customer may have to explore multiple possibilities.⁴⁶⁷ Family records, obituaries, and other components of traditional genealogical research will contribute to the proper construction of the family tree.

Someone who wants to maximize the chance of finding relatives can repeat the entire process with another DTC-GT company, but a more efficient strategy is to download the file listing the individual's SNPs and to submit it to other database companies that are willing to accept data files that they did not

460. See Greytak et al., *supra* note 457, at 107.

461. On average, a 1 centimorgan (cM) haploblock is about 1 million base pairs long, but a centimorgan is not a fixed physical distance. One cM equates to a 1% probability of recombination within the haploblock when the egg and sperm cells are formed (see *supra* section titled “Sexual Reproduction and the Genome”). Because recombination occurs more often in certain places along the genome than others, the physical distance depends on the location within the genome. Also, because recombination during the formation of egg cells happens more frequently than in the formation of sperm cells, the physical distance also differs between males and females. The randomness of the points at which “crossing over” of chromosomes occurs during meiosis gives rise to the relationship between total haploblock length and types of kinship. For a more complete explanation, see Greytak et al., *supra* note 457, at 104–06.

462. Claire L. Glynn, *Bridging Disciplines to Form a New One: The Emergence of Forensic Genetic Genealogy*, 13 *Genes* 1381, at 4 (2022), <https://doi.org/10.3390/genes13081381>; Brenna M. Henn et al., *Cryptic Distant Relatives Are Common in Both Isolated and Cosmopolitan Genetic Samples*, 7 *PLoS One* e34267 (2012), <https://doi.org/10.1371/journal.pone.0034267> (second to ninth degree cousins); Chad Huff et al., *Maximum-Likelihood Estimation of Recent Shared Ancestry (ERSA)*, 21 *Genome Res.* 768 (2011), <https://doi.org/10.1101/gr.11597> (third and even fourth degree cousins). However, “the amount of shared DNA can vary greatly for relatives of the same degree,” and “~10% of third cousins and ~50% of fourth cousins share no detectable IBD segments.” Greytak et al., *supra* note 457, at 106.

463. Ge & Budowle, *supra* note 444, at 434.

464. Greytak et al., *supra* note 457, at 107. A given individual has a first-degree relationship with his or her parents, full siblings, and children; a second-degree relationship with grandparents, aunts and uncles, half siblings, nieces and nephews, and grandchildren; a third-degree relationship with great-grandparents, great-uncles and great-aunts, first cousins, half-nieces and half-nephews, grand-nieces and grand-nephews, and great-grandchildren. For a chart, see *id.* at 105 (fig. 1).

465. *Id.* at 106 (tbl. 3) & 107; Glynn, *supra* note 462, at 5–6.

466. Greytak et al., *supra* note 457, at 106.

467. *Id.* at 107. Additional complications can arise from the history of mating in some groups. *Id.* (noting endogamy and “pedigree collapse”).

produce. Indeed, one genealogy resource, known as GEDmatch,⁴⁶⁸ does no DNA testing but lets anyone upload a SNP genotype file prepared by any major DTC-GT company. It will then perform an autosomal-haploblock comparison. The individual user receives, at no charge, information on total haploblock sharing with other GEDmatch users and whatever name (which could be an alias) and email address (which could be a one-time-use address) the possible relatives provided. In all these situations, the genealogy enthusiast never sees the raw SNP data file of anyone else.

How does IGG work?

IGG applies the same type of kinship searching to DNA recovered from a crime scene or victim to link an individual to the commission of the crime, or from human remains to identify a missing person.⁴⁶⁹ Also called forensic investigative genetic genealogy (FIGG), IGG differs from familial searching (see section above titled “Kinship trawling (‘familial searching’)”). Familial searching operates within law-enforcement-managed databases and compares STR profiles derived from legally compelled DNA samples. The grasp of this autosomal STR-based kinship trawling is basically limited to first-degree relatives of convicted offenders or arrestees. IGG is another type of “indirect” or “outer-directed” genetic search for individuals who deposited DNA associated with crimes but whose DNA is not in the accessed database. But IGG uses different DNA variants, different software, and different databases. The genealogy databases are populated with SNP DNA profiles that have been voluntarily made available to commercial database vendors for genealogical purposes. Through individually declared consent and privacy options, users of these databases can specify whether they are willing to accept law-enforcement trawling of their information. Unlike familial searching, IGG can support kinship associations beyond first-degree relatives, occasionally looking as far as ninth-degree relatives. As such, policies and legal conclusions reached for familial searching do not necessarily carry over to IGG.

IGG requires three steps to generate leads for police to consider. To begin with, as with all DNA methods, trace-evidence or victim or missing-person

468. A genetic-technology company, Verogen, specializing in forensic-sequencing methods, acquired GEDmatch in 2019, and a larger biotechnology company, Qiagen, acquired Verogen in 2023. Press Release, QIAGEN Completes Acquisition of Verogen, Strengthening Leadership in Human ID/Forensics with NGS Technologies, Jan. 9, 2023, <https://perma.cc/8UCR-AEKV>.

469. For a comprehensive and detailed review, see Daniel Kling et al., *Investigative Genetic Genealogy: Current Methods, Knowledge and Practice*, 52 *Forensic Sci. Int’l: Genetics* 102474 (2021), <https://doi.org/10.1016/j.fsigen.2021.102474>.

DNA must be recovered. This DNA normally will have been analyzed to yield an STR profile suitable for comparison with a known suspect's DNA or for trawling a CODIS database. If conventional STR profiling fails, or if queries within CODIS do not yield any matches or are not feasible because of sample limitations,⁴⁷⁰ a more laborious genetic genealogy investigation might proceed. For such an investigation, police agencies often turn to commercial forensic-science service providers, although publicly funded forensic-science laboratories are developing these capabilities.⁴⁷¹

From the forensic sample to the SNP genotype. The participating laboratory could analyze the DNA in the forensic sample with a SNP chip, as is done for the personal genealogy samples discussed in the preceding section, but microarrays may be less effective with lower quantity, degraded, or contaminated forensic samples.⁴⁷² To cope with such samples, massively parallel sequencing methods such as whole-genome or targeted sequencing⁴⁷³ have been used to identify upward of thousands to millions of SNPs and pull out a set that are either in the DTC-GT microarrays or are compatible with the databases.⁴⁷⁴ With a data file from the forensic DNA, the investigation proceeds in essentially the same manner as the usual personal genetic genealogy search.

470. When the quantity of DNA in the forensic sample is very small, and it is doubtful that STR profiling will succeed, rather than waste the DNA, a SNP profile may be generated without first trying STR profiling. In an urgent case, with enough forensic DNA, simultaneously pursuing both STR and haploblock strategies could be attractive.

471. Press Release, Univ. N. Tex. Health Sci. Cntr. at Fort Worth, CHI Becomes Nation's First Public Lab to Earn Accreditation to Perform Forensic Genetic Genealogy (Oct. 12, 2023), <https://perma.cc/F9PW-U9PQ>.

472. Kling et al., *supra* note 469, § 7.1. *But see* Greytak et al., *supra* note 457, at 103 (reporting success with trace-evidence samples of as little as one nanogram of DNA).

473. *See supra* section titled "Sequencing and SNPs."

474. Ge & Budowle, *supra* note 444, at 433; Kling et al., *supra* note 469, § 7.2. A targeted-sequencing kit that generates a smaller data file of about 10,000 widely spaced SNPs with no known clinically meaningful associations has been developed for use with GEDmatch. June Snedecor et al., *Fast and Accurate Kinship Estimation Using Sparse SNPs in Relatively Large Database Searches*, 61 *Forensic Sci. Int'l: Genetics* 102769 (2022), <https://doi.org/10.1016/j.fsigen.2022.102769>; Jessica Watson et al., *Operationalisation of the ForenSeq® Kintelligence Kit for Australian Unidentified and Missing Persons Casework*, 68 *Forensic Sci. Int'l: Genetics* 102972 (2024), <https://doi.org/10.1016/j.fsigen.2023.102972>; *see also* Andreas Tillmar et al., *The FORCE Panel: An All-in-One SNP Marker Set for Confirming Investigative Genetic Genealogy Leads and for General Forensic Applications*, 12 *Genes (Basel)* 1968 (2021), <https://doi.org/10.3390/genes12121968> (targeted sequencing kit said to include "3931 autosomal SNPs for extended kinship analysis" and "designed to exclude clinically relevant markers").

From the SNP genotype to possible relatives. The IGG service provider uploads the SNP genotype to the website of a company that maintains a database of SNP profiles for individuals looking for possible relatives (and that is willing to accept data files that it did not generate). Most if not all of these companies have procedures that allow customers to affirmatively opt in or opt out of warrantless forensic IGG trawls.⁴⁷⁵ The company's software performs haploblock matching within its database. The matching involves only the length of shared haploblocks;⁴⁷⁶ there is no need for the analyst to look at which SNPs are present within these or any other haploblocks in the forensic sample.⁴⁷⁷ The database company then informs the investigators of the names or usernames of individuals with enough shared haploblocks to merit attention as possible relatives. The total extent and pattern of the overlap in the SNP genotypes among the "matches" will be available, along with a way to contact the possible relatives. If their identities can be ascertained, traditional genealogy work begins. A genealogist constructs family trees (typically with hundreds of entries for a single third-cousin-level match), usually by scouring public records, newspapers, social media, and so on.⁴⁷⁸ The goal is to find a common ancestor for a SNP profile in the genealogy database and the forensic-sample profile. Descendants of that ancestor will include the unknown DNA source, and a combination of the demographic and genetic information may enable a genealogist or crime analyst with genealogy training to whittle down the list of descendants somewhat.⁴⁷⁹ The police now have their investigative lead. Often, it is followed by surreptitiously collecting DNA from the individuals in question to ascertain if any of them have the same STR profile as that of the forensic DNA.⁴⁸⁰

Does IGG violate the Fourth Amendment?

A series of Fourth Amendment questions can arise in connection with IGG. At the outset, one might ask whether the steps involved in IGG are a "search or

475. Heather Murphy, *What You're Unwrapping When You Get a DNA Test for Christmas*, N.Y. Times, Dec. 22, 2019.

476. See *supra* section titled "How does direct-to-consumer genetic genealogy work?"

477. Such data normally would exist only in cases in which the SNP profile was derived from more complete genome sequence data.

478. Greytak et al., *supra* note 460, at 108.

479. *Id.* at 107–10; Ge & Budowle, *supra* note 444, at 436 ("Most genealogy investigations focus on third degree or closer relationships because of the substantial burden of searching fourth degree or more matches. An efficient way to improve public record searching is triangulation, which identifies a source by intersecting between two potential families.").

480. Surreptitious DNA sampling is discussed further in the next two sections. In cases in which an STR profile cannot be obtained (see *supra* note 470), SNP-identification profiles can be used for the direct identification.

seizure” of “persons, houses, papers [or] effects.”⁴⁸¹ If they are, a further question is whether taking any or all of those steps without a warrant would violate the Amendment’s protection against “unreasonable searches and seizures.” And finally, if police apply for a warrant, what would be necessary to establish probable cause?

Are parts of IGG a search or seizure? As explained in the previous section, IGG begins with the collection of biological material from a crime scene or a victim’s remains and the extraction of DNA. To demonstrate a search or seizure, defendants would have to show either (1) that the forensic-evidence collection and preparation interferes with their right to possess or enjoy their property or personal effects⁴⁸² or (2) that simply taking possession of the stain, fluid, or cells and extracting DNA reveals information in which they have a reasonable or legitimate expectation of privacy.⁴⁸³ Few, if any, defendants have advanced such claims about biological trace evidence.⁴⁸⁴

The next step in the process is the extensive SNP analysis of the forensic sample. Is this laboratory analysis of an unknown person’s DNA a search? A number of defendants have challenged the taking—and subsequent STR genotyping—of DNA left behind by suspects in police stations or in public places.⁴⁸⁵ By and large, courts have been persuaded that STR profiling as

481. U.S. Const. amend. IV.

482. See *United States v. Jacobsen*, 466 U.S. 109, 113 (1984) (“‘seizure’ of property occurs when there is some meaningful interference with an individual’s possessory interests in that property”). Unlike the collection of DNA directly from the person, acquiring DNA from the crime scene or the remains does not entail a bodily intrusion.

483. *Katz v. United States*, 389 U.S. 347, 353 (1967); *Rakas v. Illinois*, 439 U.S. 128, 143 (1978).

484. Cf. *State v. Hartman*, 534 P.3d 423, 431–32 (Wash. Ct. App. 2023) (rejecting defendant’s argument that under a provision of the state’s constitution guaranteeing that “[n]o person shall be disturbed in his private affairs . . . without authority of law,” he “retained a privacy interest” in the DNA extracted from semen on the body of a 12-year-old girl who was found raped and murdered).

485. E.g., *United States v. Hicks*, No. 2:18-cr-20406-JTF-tmp, 2020 WL 7704556 (W.D. Tenn. May 27, 2020) (noting cases); *State v. Burns*, 988 N.W.2d 352 (Iowa 2023); cf. *State v. Athan*, 158 P.3d 27 (Wash. 2007) (detectives, posing as a fictitious law firm, sent the defendant a letter inviting him to join a fictitious class-action lawsuit, to obtain DNA from the return envelope to compare to DNA from the crime scene). Commentators have questioned the “abandonment” line of cases. E.g., Elizabeth E. Joh, *Reclaiming “Abandoned” DNA: The Fourth Amendment and Genetic Privacy*, 100 Nw. U. L. Rev. 857–84 (2006); Tracey Maclin, *Government Analysis of Shed DNA Is a Search Under the Fourth Amendment*, 48 Tex. Tech. L. Rev. 287 (2015); Albert E. Scherr, *Genetic Privacy and the Fourth Amendment: Unregulated Surreptitious DNA Harvesting*, 47 Ga. L. Rev. 445 (2013). But see Edward J. Imwinkelried & David H. Kaye, *DNA Typing: Emerging or Neglected Issues*, 76 Wash. L. Rev. 413, 437 (2001); David H. Kaye, *Science Fiction and Shed DNA*, 101 Nw. U. L. Rev. Colloquy 62 (2006), <https://perma.cc/596A-NKH7> (response to Joh).

conducted in the laboratory solely to produce identifying information from “abandoned” biological material is no more invasive of privacy than examining fingerprints from a crime scene.⁴⁸⁶ Singly, or more likely in combination, however, many of the SNPs in the raw data file could serve as markers or predictors of disease.⁴⁸⁷ Yet, neither the police nor the laboratories working for them inspect the SNP data files for such markers to make haploblock comparisons.⁴⁸⁸ Is this fact sufficient to defeat the claim that the very possession of a list of SNPs in a text file interferes with a reasonable expectation of privacy (even before the identity of the individual whose DNA has been analyzed is known)?⁴⁸⁹ If not, would legally enforceable rules against misuse or unauthorized disclosure create a convincing analogy to photographs, fingerprints, or CODIS STR profiles (assuming that the act of comparing such biometric identifiers is not itself a search)? Or is the creation of the SNP data file to find the source of the as-yet-unidentified DNA more like the actual perusal of the text and image files in smartphones⁴⁹⁰ or the acquisition of cell-site location information on personal movements over long time periods⁴⁹¹—both Fourth Amendment searches?

The next step in IGG is the haploblock-sharing trawl. Is the trawl performed by the company that operates the commercial genealogy database a governmental search⁴⁹² of the defendant’s person or effects? A series of

486. Opinions zeroing in on this issue include *United States v. Davis*, 690 F.3d 226, 243–46 (4th Cir. 2012) (separate search); *Burns*, 988 N.W.2d at 364–65 (no search); *Raynor v. State*, 99 A.3d 753 (Md. 2014) (no search); *Commonwealth v. Arzola*, 26 N.E.3d 185, 190–94 (Mass. 2015) (no search). For academic commentary on STR profiling as a search unto itself, see, for example, Kaye, *supra* note 445; Maclin, *supra* note 485; Murphy, *supra* note 446.

487. To reduce this risk, sequencing kits that generate smaller data files of widely spaced SNPs with no known clinically meaningful associations have been developed. See *supra* note 474.

488. Ellen M. Greytak et al., *Privacy and Genetic Genealogy Data*, 361 Science 857 (2018), <https://doi.org/10.1126/science.aav0330>; Greytak et al., *supra* note 457, at 106–07 (“Additionally, no sensitive genetic information is disclosed to law enforcement during a genetic genealogy search, as the raw genetic data from GEDmatch users is not accessible . . . GEDmatch simply performs comparisons among samples, returning the lengths and chromosomal locations of shared DNA segments, which are used to determine the approximate relationship between individuals.”).

489. Initially, the investigators’ SNP data file does not pertain to any known individual, but if IGG is successful and the source of the trace-evidence sample is located, then the data will pertain to that person.

490. *Riley v. California*, 573 U.S. 373 (2014) (viewing data stored on a cellphone is a search that does not fall within the search-incident-to-arrest exception to the warrant-and-probable-cause requirement).

491. *Carpenter v. United States*, 585 U.S. 296 (2018).

492. On the question of whether ordinary, non-kinship trawls in law-enforcement STR databases are Fourth Amendment “searches,” see *Boroian v. Mueller*, 616 F.3d 60 (1st Cir. 2010) (retrawling after a sentence is served is not a search); *Johnson v. Quander*, 440 F.3d 489, 498–99 (D.C. Cir. 2006) (retrawling after the period of probation is completed is not a search); David H. Kaye, *DNA Database Trawls and the Definition of a Search* in *Boroian v. Mueller*, 97 Va. L. Rev. in Brief 41 (2011);

subquestions may need to be considered. Did the company knowingly and voluntarily trawl the database to assist the investigation?⁴⁹³ Or did the company not intend to make its services available to law enforcement,⁴⁹⁴ but police submitted SNP data files anyway, pretending to be an ordinary subscriber? Would a customer, having chosen to expose his or her SNP data to trawling on behalf of anyone in the general public, retain a reasonable expectation of privacy in that data?⁴⁹⁵ And, even if the trawl were considered a search of the customer's data, could a defendant suppress the evidence on the theory that the

Catherine W. Kimel, Note, *DNA Profiles, Computer Searches, and the Fourth Amendment*, 62 Duke L.J. 933 (2013).

493. FamilyTreeDNA and GEDmatch require law-enforcement agencies to identify themselves as such when submitting a SNP-genotype file. FamilyTreeDNA Law Enforcement Guide, <https://perma.cc/PHR5-2B59>, last visited, Dec. 16, 2023; GEDmatch, Terms of Service and Privacy Policy, Dec. 30, 2021, <https://perma.cc/FJ5G-AL6F>.

494. *E.g.*, MyHeritage—Terms and Conditions (June 11, 2023), <https://perma.cc/DJP9-D7UZ> (“using the DNA Services for law enforcement purposes, forensic examinations, criminal investigations, ‘cold case’ investigations, identification of unknown deceased people, location of relatives of deceased people using cadaver DNA, and/or all similar purposes, is strictly prohibited, unless a court order is obtained”); GEDmatch, Terms of Service and Privacy Policy (Dec. 30, 2021), <https://perma.cc/LF67-BBKD> (if you “[o]pt-out . . . [y]our kit WILL NOT be compared with kits submitted by law enforcement to identify perpetrators of violent crimes”) (emphasis in original).

495. It could be argued that just because DNA samples are sent to DTC-GT companies for SNP analysis, or because the SNP data are made available to the public for genetic-genealogy trawls, no cognizable search occurs when the government also submits a data file for trawling and receives the results. *Cf., e.g.*, *United States v. Miller*, 425 U.S. 435, 443 (1976) (“The [customer who maintains a bank account] takes the risk, in revealing his affairs to another, that the information will be conveyed by that person to the Government. . . . [T]he Fourth Amendment does not prohibit the obtaining of information revealed to a third party and conveyed by him to Government authorities, even if the information is revealed on the assumption that it will be used only for a limited purpose and the confidence placed in the third party will not be betrayed.”). In *Carpenter v. United States*, 585 U.S. 296 (2018), however, the Court clarified that this “third-party doctrine” depends on the nature of the information compiled by or disclosed to third parties. It held that extensive data on the locations of a person's cell phone relative to cell towers reveals too much about a person's private life to be accessible to police at almost no cost just because the person's movements “might be pertinent to an ongoing investigation.” *Id.* at 317. *Carpenter* opens the door for subscribers to claim that the genealogy database trawl is a Fourth Amendment “search” of their data. They can argue that exposing them to the risk of being placed on a family tree known to the government is no less invasive of a legitimate privacy interest than a cellphone customer “effectively [being] tailed every moment of every day for five years.” *Id.* at 312. But there is public and scholarly disagreement on the strength of the interests of the subscribers to genetic genealogy services in the secrecy of their biological affiliations and their SNPs. Compare James W. Hazel & Christopher Slobogin, “A World of Difference”? *Law Enforcement, Genetic Data, and the Fourth Amendment*, 70 Duke L.J. 705 (2021); Ayesha K. Rasheed, *Personal Genetic Testing and the Fourth Amendment*, 2020 U. Ill. L. Rev. 1249; and Natalie Ram, *Genetic Privacy After Carpenter*, 105 Va. L. Rev. 1357 (2019); with Teneille R. Brown, *Why We Fear Genetic Informants: Using Genetic Genealogy To Catch Serial Killers*, 21 Colum. Sci. & Tech. L. Rev. 114 (2019); cf. Paul Ohm, *The Many Revolutions of Carpenter*, 32 Harv. J.L. & Tech. 357, 385 (2019) (concluding that even for a database of whole-genome

use of the customer's data unreasonably invaded the customer's Fourth Amendment interests?⁴⁹⁶

Warrants and probable cause? If some or all of these steps in IGG are individually or collectively a Fourth Amendment search of the defendant's person, papers, or effects, two broad questions remain: first, whether such searches are constitutionally unreasonable in the absence of a warrant based on probable cause; and, second, if a warrant is sought, what must be shown to establish probable cause. With regard to reasonableness, a few Supreme Court opinions suggest that "our general Fourth Amendment approach [entails] 'examining the totality of the circumstances. . . .'"⁴⁹⁷ In IGG cases, parties might argue about the relevance and impact of many circumstances. For example, did anyone actually scrutinize the defendant's SNP profile for medically or socially sensitive information? Are there statutory, administrative, or technical constraints against using the SNP data for such purposes?⁴⁹⁸ Is the system designed to avoid unnecessary disclosures of family information, such as misattributed paternity? Is the prospect that law-enforcement authorities will initiate trawls of the database known to the subscribers?⁴⁹⁹ These privacy-related factors might affect the magnitude of the individual interests as compared to government's law-enforcement interest.

But directly balancing the two sets of interests to determine the reasonableness of a Fourth Amendment search or seizure is controversial. More often, the Supreme Court has admonished that "searches conducted outside the judicial

sequences "[i]t seems unlikely that a court would require a warrant for DNA evidence held by a private third party based on a straight application of the *Carpenter* factors").

496. The defendant would have to overcome the doctrine that an impermissible intrusion on the security or privacy of another person's papers or effects does not entitle the defendant to suppress the evidence. See, e.g., *Rakas v. Illinois*, 439 U.S. 128, 133–34 (1978) ("A person who is aggrieved by an illegal search and seizure only through the introduction of damaging evidence secured by a search of a third person's premises or property has not had any of his Fourth Amendment rights infringed."); *Minnesota v. Carter*, 525 U.S. 83, 91 (1998) (a temporary guest in an apartment does not have standing to challenge the reasonableness of a search of the apartment).

497. *United States v. Knights*, 534 U.S. 112, 118 (2001). With CODIS STR databases, lower courts using this "totality" approach typically balanced the government's interests in building and operating these intelligence-gathering databases against the individual's legitimate interests in being free from bodily intrusion and maintaining the confidentiality of the STR data and samples. See *supra* section titled "Law-enforcement databases and databanks."

498. Cf. *Maryland v. King*, 569 U.S. 435 (2013) (DNA sampling and CODIS STR profiling of arrestees); *Whalen v. Roe*, 429 U.S. 589 (1977) (law requiring that a copy of prescriptions for the most dangerous legitimate drugs be sent to the state for computer scanning, storage, and access under secure conditions was a reasonable exercise of state police power).

499. See *supra* section titled "From the SNP genotype to possible relatives" (on policies allowing customers to opt in or opt out of exposure to database trawls for law-enforcement purposes).

process, without prior approval by judge or magistrate, are per se unreasonable under the Fourth Amendment—subject only to a few specifically established and well-delineated exceptions.”⁵⁰⁰ Under this “warrant preference rule,” balancing to determine the reasonableness of a type of search that dispenses with warrants or probable cause still might occur, but only at the level of delineating a new categorical exception to these procedures.⁵⁰¹ For a warrant to issue, a magistrate would need to verify that the police have probable cause. But probable cause to believe what? The Supreme Court has said that probable cause to arrest or search a person exists when the totality of the circumstances indicates “reasonable ground for belief of guilt,” and “the belief of guilt [is] particularized with respect to the person to be searched or seized.”⁵⁰² Genealogy database trawls are not undertaken because of a belief about a particular person. They are initiated in the hope that some individuals unknown to the police have their SNP genotypes in the database and that the company’s haploblock-matching software will flag them as possible relatives. Should the probable-cause finding turn on whether the size and likely composition of a company’s database indicates a reasonable probability of a useful match occurring in that particular database? Why not all databases combined, since the trawls could be performed in all of them? Or are such warrants too redolent of “the reviled ‘general warrants’ and ‘writs of assistance’ of the colonial era, which allowed British officers to rummage through homes in an unrestrained search for evidence of criminal activity”?⁵⁰³ The nature of probable cause in the context of IGG may need to be explored.

How should IGG be regulated?

After IGG burst into the public consciousness with the arrest in 2018 of a former police officer found to be the elusive Golden State Killer,⁵⁰⁴ DTC-GT companies scrambled to develop clearer or different policies,⁵⁰⁵ and observers advocated

500. *Katz v. United States*, 389 U.S. 347, 357 (1967) (notes omitted).

501. David H. Kaye, *A Fourth Amendment Theory for Arrestee DNA and Other Biometric Databases*, 15 U. Pa. J. Const. L. 1095, 1139–40 (2013) (proposing a “biometric exception” that would permit warrantless searches of fingerprint, photograph, and, arguably, CODIS STR databases). For efforts to explain the rationale for the balancing for DNA sampling that the Court performed in *Maryland v. King*, see, for example, Kaye, 104 J. Crim. L. & Criminology 535, *supra* note 431; Maclin, *supra* note 431.

502. *Maryland v. Pringle*, 540 U.S. 366, 371 (2003) (internal quotation marks and citations omitted).

503. *Riley v. California*, 573 U.S. 373 (2014).

504. See, e.g., Paige St. John, *The Untold Story of How the Golden State Killer Was Found: A Covert Operation and Private DNA*, L.A. Times, Dec. 8, 2020.

505. Glynn, *supra* note 462, at 10; Murphy, *supra* note 475.

for external rules and oversight of when and how police access the databases.⁵⁰⁶ A year and a half later, the Department of Justice settled on an “interim policy” for its agencies, employees, and contractors.⁵⁰⁷ This policy

- Mostly limits the use of IGG to unsolved homicides and sex crimes.⁵⁰⁸
- Restricts investigators to “only those GG services that provide explicit notice to their service users and the public that law enforcement may use their service sites.”⁵⁰⁹
- Requires that “biological samples and [IGG] profiles [be used] only for law enforcement identification purposes”⁵¹⁰ and not “to determine the sample donor’s genetic predisposition for disease or any other medical condition or psychological trait.”⁵¹¹
- Establishes procedural constraints on surreptitious collection of DNA from certain individuals who are not targets of the investigation.⁵¹²
- Requires that all forensic samples first be tested to generate an STR profile prior to performing IGG, a practice that is not feasible with low input, degraded, and rootless hair samples.

506. Sara H. Katsanis, *Pedigrees and Perpetrators: Uses of DNA and Genealogy in Forensic Investigations*, 21 Ann. Rev. Genomics & Hum. Genetics 535 (2020), <https://doi.org/10.1146/annurev-genom-111819-084213>; Erin Murphy, *Law and Policy Oversight of Familial Searches in Recreational Genealogy Databases*, 292 Forensic Sci. Int’l e5–e9 (2018), <https://doi.org/10.1016/j.forsciint.2018.08.027>; Paige St. John, *DNA Genealogical Databases Are a Gold Mine for Police, But with Few Rules and Little Transparency*, L.A. Times, Nov. 24, 2019.

507. U.S. Dep’t of Just., Interim Policy Forensic Genetic Genealogical DNA Analysis and Searching, Nov. 1, 2019, <https://perma.cc/D6ZF-D6SW>. The policy also applies to “any federal agency or any unit of state, local, or tribal government that receives grant award funding from the Department that is used to conduct FGG/FGGS.” *Id.* at 2. However, the policy “does not . . . limit the prerogatives, choices, or decisions available to, or made by, the Department in its discretion.” *Id.* at 1 n.1.

508. These include missing-person cases where the unidentified human remains are reasonably believed to be those of a suspected homicide victim. *Id.* at 4. However, IGG can be used for other crimes when (1) they are “serious” and the database vendor authorizes “investigative use of its service by law enforcement,” *id.* at 4 n.15, or (2) “the circumstances surrounding the criminal act(s) present a substantial and ongoing threat to public safety or national security.” *Id.* at 5. Reasonable investigative leads must have been pursued, and a CODIS search must have failed. *Id.*

509. *Id.* at 6.

510. *Id.*

511. *Id.* at 7.

512. The protected individuals are non-suspects “who may have a closer kinship relationship to the donor of the forensic sample than the associated [genetic genealogy] service user” and who have come to light as a result of the genealogical research. *Id.* at 6. In such cases, investigators either must obtain their informed consent or have “reasonable grounds to believe that this request would compromise the integrity of the investigation” and consult prosecutors before covertly collecting their samples. *Id.* In addition, investigators must obtain a search warrant for “a vendor laboratory” to perform SNP typing on the sample. *Id.*

- Requires that only STR profiles be used for confirmatory identity verification against a sample collected from the putative perpetrator or missing person, when SNP-based statistics might more easily be generated with certain samples.
- Creates other constraints and procedural requirements as well.

In sum, the interim policy stakes out a middle ground between unfettered IGG and a total ban. A few states have enacted more restrictive statutes that include judicial oversight.⁵¹³ Both the Justice Department policy and the statutes have their critics.⁵¹⁴

Inferring Traits

Kinship searching (see section titled “Kinship trawling (‘familial searching’) above”) is one technological response to unsuccessful database searches. Two additional techniques are available: biogeographic ancestry testing and forensic DNA phenotyping. Their purpose is only to focus an investigation by supplying police with some visible or other characteristics that the source of the trace evidence might have. They are not routinely used and are not designed to produce evidence for trial.⁵¹⁵ If a suspect is located with the help of these probabilistic clues, STR profiling (or typing with other loci suitable for individual identification) would be undertaken before trial.⁵¹⁶

513. *E.g.*, Utah Code § 53-10-403.7 (2023); Natalie Ram et al., *Regulating Forensic Genetic Genealogy: Maryland’s New Law Provides a Model for Others*, 373 Science 1444 (2021), <https://doi.org/10.1126/science.abj5724>.

514. See Virginia Hughes, *Two New Laws Restrict Police Use of DNA Search Method: Maryland and Montana Have Passed the Nation’s First Laws Limiting Forensic Genealogy, the Method that Found the Golden State Killer*, N.Y. Times, May 31, 2021; Laura Geary, *A Critical Eye Toward Commercial DNA Database Criminal Procedures*, U. Chi. L. Rev. Online, July 8, 2022, <https://perma.cc/TG3D-U2GZ> (questioning the statutes); Christi Guerrini et al., *State Genetic Privacy Statutes: Good Intentions, Unintended Consequences?*, Bill of Health (Aug. 10, 2023), <https://perma.cc/2VGQ-S4L7> (pointing out ambiguities and anomalies); *cf.* Brown, *supra* note 495 (critiquing some of the concerns about IGG); Christi J. Guerrini et al., *Four Misconceptions About Investigative Genetic Genealogy*, 8 J. L. & Biosciences 1 (2021), <https://doi.org/10.1093/jlb/lsab001> (urging more nuanced analysis of the ethical and practical issues).

515. *But see* Jonathan Foox et al., *Epigenetic Forensics for Suspect Identification and Age Prediction*, 1 Forensic Genomics 83 (2021), <https://doi.org/10.1089/forensic.2021.0005> (experiment with DNA-based age estimation offered to rebut offender’s claim that an old DNA sample from a previous case was planted in the newer case for which post-conviction relief was denied).

516. Indeed, police have used the DNA-based information in conducting “mass screens” or “DNA dragnets” of legally consenting residents of a geographic region where serial rapes, murders, or bombings have occurred. *E.g.*, *Monroe v. City of Charlottesville*, 579 F.3d 380 (4th Cir. 2009). There is a risk that a screening will be restricted on the basis of race or ethnicity to the wrong group. But DNA predictions also could prompt police to change a screen that otherwise would be

Biogeographic ancestry testing (BGAT)

Some DNA sequences are concentrated in some parts of the world. As such, a well-chosen set of loci can be used to infer the geographic area in which an individual's biological ancestors lived. The specificity and accuracy of these inferences depends on several factors. What is the size and scale of the population-genetics databases used to associate genotypes with geographic regions? How clearly do the alleles in the panel of loci cluster according to geography (or self-declared ancestry, if that is used)? How effective is the statistical method for probabilistically mapping genotypes to geography?

These matters are all subjects of ongoing research. As potential ancestry-informative markers (AIMs), the STR loci used for individual identification are far from ideal. They vary greatly among individuals—every population has many alleles at each locus—but it is roughly the same “many” in most populations. For ancestry assignment, various panels of SNPs that are more strongly correlated to geography have been published.⁵¹⁷ Which ancestry-informative SNPs are present in a trace-evidence sample can be determined with most of the methods described above in the section titled “What Are DNA Polymorphisms and How Are They Detected?”⁵¹⁸ Whereas clinical and anthropological research studies often use thousands of SNPs, the panels developed for biological trace evidence, which frequently involves low quantity or degraded DNA, consist of smaller sets of AIMs.⁵¹⁹

incorrectly concentrated on one group. Imwinkelried & Kaye, *supra* note 485, at 451. On the equal-protection ramifications of racially focused investigations resulting from eyewitness statements, see Monroe, *supra*; Brown v. City of Oneonta, 221 F.3d 329 (2d Cir. 2000); Imwinkelried & Kaye, *supra* note 485, at 446–51.

517. For a review, see Ozlem Bulbul & Kenneth K. Kidd, *Forensic Ancestry Inference: Data Requirements, Analysis Methods, and Interpretation of Results*, in *Forensic DNA Analysis: Technological Development and Innovative Applications* 225, 227–29 (Elena Pilli & Andrea Berti eds., 2021).

518. *Id.* at 229–31; Elena Pilli et al., *Biogeographical Ancestry, Variable Selection, and PLS-DA Method: A New Panel to Assess Ancestry in Forensic Samples via MPS Technology*, 62 *Forensic Sci. Int'l: Genetics* 102806 (2023), <https://doi.org/10.1016/j.fsigen.2022.102806>.

519. Peter Resutik et al., *Comparative Evaluation of the MAPlex, Precision ID Ancestry Panel, and VISAGE Basic Tool for Biogeographical Ancestry Inference*, 64 *Forensic Sci. Int'l: Genetics* 102850 (2023), <https://doi.org/10.1016/j.fsigen.2023.102850> (“More than twenty forensic panels designed to distinguish continental ancestry have been proposed by the forensic community over the past two decades. These generally comprise less than 200 AIMs, mainly consisting of autosomal SNPs, but also including other types of markers such as microhaplotypes (MHs) and insertion-deletion polymorphisms (INDELs). In addition, the inclusion of uniparental markers (Y chromosomal SNPs and mtDNA) in the same panel or as a complementary approach has been advocated as it enables ascertaining information on the maternal and paternal lineages, which may help resolve ancestry of recently-admixed individuals.”) (references omitted).

The population-reference databases for developing the ancestry-informative SNPs for crime-scene samples are public research databases.⁵²⁰ Their limited coverage of populations around the world constrains the inferred origins to very large areas. Basically, forensic BGAT is most reliable for mapping a sample to a major continental group such as African, Western Eurasian, East and South Asian, or Native American. Because significant numbers of SNPs are employed as markers, established multivariate statistical methods are used to make the classifications.⁵²¹ Studies of the accuracy of classifications from different providers continue to appear.⁵²²

Geneticists and anthropologists typically caution that BGA should not be confused with ethnicity and race. Ethnicity reflects a person's social and cultural background. Obviously, these are not determined by DNA. However, a person's ethnicity may be strongly associated with BGA. As a biological classification, the term "race" is considered inappropriate, but self-reported race also is correlated with BGA.⁵²³

Forensic DNA phenotyping (FDP)

A phenotype is an observable or testable trait—skin color, height, blood sugar level, diseases, physical activity level, etc. Phenotypes result from gene activity and environmental factors. Forensic DNA phenotyping refers to inferring traits from the DNA in a trace-evidence sample to assist in locating the source. As such, it might be clearer to call it DNA trace-evidence phenotyping. Regardless of nomenclature, researchers and practitioners have focused on deducing

520. The 1000 Genomes Project, Human Genome Diversity Project-Center d'Étude du Polymorphisme Humain (HGDP-CEPH), and HapMap Project databases are commonly used. A short description of their coverage, along with that of the ALFRED database of frequency of polymorphisms in human populations, is at Bulbul & Kidd, *supra* note 517, at 231–32.

521. *Id.* at 232–37. Machine-learning methods also are being studied. Eugenio Alladio et al., *Multivariate Statistical Approach and Machine Learning for the Evaluation of Biogeographical Ancestry Inference in the Forensic Field*, 12 *Sci. Reports* 8974 (2022), <https://doi.org/10.1038/s41598-022-12903-0>.

522. Muna Al-Asfi et al., *Assessment of the Precision ID Ancestry Panel*, 132 *Int'l J. Legal Med.* 1581 (2018), <https://doi.org/10.1007/s00414-018-1785-9>; Lauren Atwood et al., *From Identification to Intelligence: An Assessment of the Suitability of Forensic DNA Phenotyping Service Providers for Use in Australian Law Enforcement Casework*, 11 *Frontiers in Genetics* 568701 (2021), <https://doi.org/10.3389/fgene.2020.568701>; Resutik et al., *supra* note 519; Watson et al., *supra* note 487.

523. For discussion of populations and race in forensic genetics, see, for example, Jobling, *supra* note 183; Mark Shriver et al., *Getting the Science and the Ethics Right in Forensic Genetics*, 37 *Nature Genetics* 449 (2005), <https://doi.org/10.1038/ng0505-449>; *supra* section titled "Could an Unrelated Person Be the Source?"; cf. Alyna T. Khan et al., *Race, Ethnicity, and Genetics Working Group, Recommendations on the Use and Reporting of Race, Ethnicity, and Ancestry in Genetic Research: Experiences from the NHLBI TOPMed Program*, 2 *Cell Genomics* 100155 (2022), <https://doi.org/10.1016/j.xgen.2022.100155>.

externally visible characteristics of the kind that might appear in an eyewitness's description.⁵²⁴

One such characteristic (or, more precisely, a state that is associated with many aspects of physical appearance) is biological sex. A locus for sex typing has been part of standard STR profiling for over 25 years.⁵²⁵ Eye color is harder to infer for all people because many genes are involved, but only six are currently used in FDP tests. Some classifications (such as brown eye color versus blue) often can be made,⁵²⁶ and, when they are possible, they seem to be generally accurate.⁵²⁷ Classifications of hair and skin pigmentation also are included in some FDP-typing systems whose validity has been studied in a number of populations.⁵²⁸ Predicting other externally visible traits such as height is even harder because these quantitative traits are determined by large numbers of genes, many of which remain unknown. These traits are also influenced by environmental factors such as nutrition, which cannot be predicted from DNA. Although some police agencies have turned to bioinformatics companies that construct pictures of faces from DNA, this procedure has been criticized as unvalidated, not peer reviewed, and beyond existing knowledge.⁵²⁹

In addition to FDP for characteristics of eyes, hair, and skin, estimation of chronological age is possible.⁵³⁰ Mitochondrial DNA changes with age (and not

524. Scientific literature on FDP is referenced in John M. Butler, *Recent Advances in Forensic Biology and Forensic DNA Typing: INTERPOL Review 2019–2022*, 6 *Forensic Sci. Int'l: Synergy* 100311 (2023), <https://doi.org/10.1016/j.fsisy.2022.100311>.

525. See *supra* note 48.

526. Eye color depends on how much of a dark brownish-black pigment called eumelanin is in the iris. Eumelanin absorbs light, so individuals who have a lot of eumelanin in their irises have eye colors that appear dark, such as brown. Several genes responsible for controlling eumelanin production in the eye are known. One of the most important (OCA2) is located on chromosome 15. Individuals with two copies of the G nucleotide at a locus near OCA2 produce less eumelanin and are more likely to have blue eyes. Individuals with one or two copies of an A nucleotide at the same SNP locus produce more eumelanin and are more likely to have brown eyes. Intermediate eye colors such as green and hazel are still very hard to predict.

527. E.g., Ersilia Paparazzo et al., *A New Approach to Broaden the Range of Eye Colour Identifiable by IrisPlex in DNA Phenotyping*, 12 *Sci. Reports* 12803 (2022), <https://doi.org/10.1038/s41598-022-17208-w> (describing the system as “well validated”).

528. See European Forensic Genetics Network of Excellence, *Making Sense of Forensic Genetics* 31 (2017), <https://perma.cc/4RSH-5EQM> (“Currently, eye colour, hair colour and skin colour can be predicted reliably and with practically useful accuracy from crime scene DNA, but not yet any other externally visible characteristic.”) (quoting Manfred Kayser); cf. Matteo Fabbri et al., *Application of Forensic DNA Phenotyping for Prediction of Eye, Hair and Skin Colour in Highly Decomposed Bodies*, 11 *Healthcare* 647 (2023), <https://doi.org/10.3390/healthcare11050647>.

529. *Id.*; Jobling, *supra* note 183.

530. Gerontology researchers often distinguish between chronological and biological age (also known as physiological or functional age). Biological aging occurs as damage to various cells and tissues in the body accumulates. Biological age reflects chronological age, genetics (for example, how quickly antioxidant defenses kick in), lifestyle, nutrition, and diseases and other

for the better),⁵³¹ but the association of these modifications with age is weak. In nuclear DNA, sequences near the ends of the chromosomes shorten with every cell division. Estimates based on this effect are not very precise (typical errors are around ± 20 years, perhaps less when using certain blood cells). Better results come from changes to DNA that do not alter the sequence of nucleotides.⁵³² Such “epigenetic” changes occur in connection with gene expression. Messenger RNA (mRNA) participates in this gene expression, and other molecules attach to the DNA in the process that turns gene expression on and off.⁵³³ In particular, a methyl group (one carbon atom and three hydrogen atoms) on top of a gene turns it off (or sometimes on). The extent and pattern of this methylation across the genome changes as an individual ages. This phenomenon enables DNA methylation and mRNA to serve as age markers. Analytical procedures for markers in large trace-evidence samples seem to give estimated ages that usually are accurate within about ± 3 or 4 years; massively parallel sequencing for methylation in samples containing smaller amounts of DNA is in development⁵³⁴ and has been used in one prominent criminal case.⁵³⁵

FDP has prompted criticism from some legal commentators and bioethicists.⁵³⁶ Proponents of the methods maintain that FDP merely reveals information for which a person leaving a trace-evidence sample could have no plausible expectation of privacy, since visible traits are exposed to the public anyway. Beyond phenotypes like pigmentation, “lifestyle factors, such as smoking, alcohol intake, drug abuse, diet, physical exercise and educational attainment have been reported to impact our epigenome” and to merit research.⁵³⁷ But some commentators regard “lifestyle and environmental characteristics” as included in “the right to privacy.”⁵³⁸ Other writers note that “if FDP techniques

conditions. Biological age is more closely connected to appearance, but chronological age can be more directly useful in investigations.

531. A higher load of point heteroplasmies and an increasing number of large-scale deletions occur.

532. Vidaki & Kayser, *supra* note 398.

533. See *supra* section titled “Genes and Gene Products.”

534. Walther Parson, *Age Estimation with DNA: From Forensic DNA Fingerprinting to Forensic (Epi)Genomics: A Mini-Review*, 64 *Gerontology* 326 (2018), <https://doi.org/10.1159/000486239>; Vidaki & Kayser, *supra* note 398.

535. *State v. Avery*, 970 N.W.2d 564 (Wis. Ct. App. 2021) (table); Foox et al., *supra* note 515.

536. Gabrielle Samuel & Barbara Prainsack, *Societal, Ethical, and Regulatory Dimensions of Forensic DNA Phenotyping* (2019), <https://perma.cc/XT53-H35S> (surveying literature and summarizing interviews); Pilar N. Ossorio, *About Face: Forensic Genetic Testing for Race and Visible Traits*, 34 *J. L., Med. & Ethics* 277 (2006), <https://doi.org/10.1111/j.1748-720X.2006.00033.x>; Victor Toom et al., *Approaching ethical, legal and social issues of emerging forensic DNA phenotyping (FDP) technologies comprehensively*, 22 *Forensic Sci. Int'l: Genetics* e1 (2016), <https://doi.org/10.1016/j.fsigen.2016.01.010>.

537. Vidaki & Kayser, *supra* note 398.

538. Matthias Wienroth et al., *Ethics as Lived Practice. Anticipatory Capacity and Ethical Decision-Making in Forensic Genetics*, 12 *Genes* 1868 (2021), <https://doi.org/10.3390/genes12121868>.

were extended to [certain] non-visible characteristics, such as genetic risk of disease, they could reveal personally sensitive, private information, to whoever is doing the testing—which could be of interest to medical insurance companies or certain employers.”⁵³⁹ In addition, there are questions of premature use of methods that have not attained adequate validation and precision, and of effectively communicating uncertainties. Also, there is concern that police investigations informed by surmises about skin color will “be seen as reinforcing racial stereotyping and perceptions of unequal treatment amongst minority communities.”⁵⁴⁰

Whether trace-evidence samples can be used for phenotyping has not been litigated. For samples collected for storage in law-enforcement databanks, most legislation establishing and funding law-enforcement DNA databases simply uses the broad term “identification” to denote the purpose behind collecting and storing samples, presumably from convicted offenders or arrestees rather than from the trace evidence. However, a few states exclude testing these samples for “physical characteristics, traits or predispositions for disease.”⁵⁴¹

539. European Forensic Genetics Network of Excellence, *supra* note 528. An outpouring of legislation forbids the use of genetic-test results in health insurance and employment. *See generally* Jean-Christophe Bélisle-Pipon et al., *Genetic Testing Insurance Discrimination and Medical Research: What the United States Can Learn from Peer Countries*, 25 *Nature Med.* 1198 (2019), <https://doi.org/10.1038/s41591-019-0534-z>; Mark A. Rothstein, *Time to End the Use of Genetic Test Results in Life Insurance Underwriting*, 46 *J.L., Med. & Ethics* 794 (2018), <https://doi.org/10.1177/1073110518804243> (“By the end of the [1990s], 48 states had enacted laws prohibiting genetic discrimination in health insurance and 35 states prohibited genetic discrimination in employment. In 2008, Congress enacted the Genetic Information Nondiscrimination Act (GINA), which outlawed discrimination based on genetic information in health insurance and employment.”) (notes omitted).

540. Weinroth et al., *supra* note 538, at 32 (quoting Robin Williams).

541. Indiana Code § 10-13-6-16 (“The information contained in the Indiana DNA data base may not be collected or stored to obtain information about human physical traits or predisposition for disease.”); Wyo. Stat. § 7-19-404(c) (2022) (“Only DNA records which directly relate to the identification characteristics of individuals shall be collected and stored in the state DNA database. The information contained in the state DNA database shall not be collected or stored for the purpose of obtaining information about physical characteristics, traits or predisposition for disease.”); *cf.* R.I. Gen. L. § 12-1.5-10(5) (2022) (“DNA samples and DNA records collected under this chapter shall never be used under the provisions of this chapter for the purpose of obtaining information about physical characteristics, traits or predispositions for disease.”). On the desirability of limiting the loci that may be typed from offender samples in law-enforcement databanks, compare Mark A. Rothstein & Sandra Carnahan, *Legal and Policy Issues in Expanding the Scope of Law Enforcement DNA Data Banks*, 67 *Brook. L. Rev.* 127 (2001), with David H. Kaye, *Two Fallacies about DNA Data Banks for Law Enforcement*, 67 *Brook. L. Rev.* 179 (2001). A few European countries have laws that affect phenotyping with trace-evidence samples. Henrik Westermarck et al., *Swiss Inst. Comparative L., The Regulation of the Use of DNA in Law Enforcement*, Aug. 28, 2020, at 10, <https://perma.cc/B3K6-DK3T>.

Glossary of Terms

adenine (A). One of the four bases, or nucleotides, that make up the DNA molecule. Adenine binds only to thymine. See nucleotide.

affinal method. A method for computing the single-locus profile probabilities for a theoretical subpopulation by adjusting the single-locus profile probability, calculated with the product rule from the mixed population database, by the amount of heterogeneity across subpopulations. The model is appropriate even if there is no database available for a particular subpopulation, and the formula always gives more conservative probabilities than the product rule applied to the same database.

allele. In classical genetics, an allele is one of several alternative forms of a gene. A biallelic gene has two variants; others have more. Alleles are inherited separately from each parent, and for a given gene an individual may have two different alleles (heterozygosity) or the same allele (homozygosity). In DNA analysis, the term is applied to any DNA region (even if it is not a gene) used for analysis.

allelic ladder. A mixture of all the common alleles at a given locus. Periodically producing electropherograms of the allelic ladder aids in designating the alleles detected in an unknown sample. The positions of the peaks for the unknown can be compared to the positions in a ladder electropherogram produced near the time when the unknown was analyzed. Peaks that do not match up with the ladder require further analysis.

amplification. Increasing the number of copies of a DNA region, usually by PCR.

amplified fragment length polymorphism (AMP-FLP). A DNA identification technique that uses PCR-amplified DNA fragments of varying lengths. The DS180 locus is a VNTR whose alleles can be detected with this technique.

ancestry informative markers (AIMs). Polymorphisms for a particular DNA sequence that appear in substantially different frequencies among populations from different geographical regions of the world. AIMs are used to infer the geographical origins of the ancestors of an individual, typically by continent or subcontinent. Panels of SNPs are usually used, as they vary more in their frequency across major populations than do the STRs used in individual identification.

antibody. A protein (immunoglobulin) molecule, produced by the immune system, that recognizes a particular foreign antigen and binds to it; if the antigen is on the surface of a cell, this binding leads to cell aggregation and subsequent destruction.

antigen. A molecule (typically found on the surface of a cell) whose shape triggers the production of antibodies that will bind to the antigen.

autoradiograph (autoradiogram, autorad). In RFLP analysis, the X-ray film (or print) showing the positions of radioactively marked fragments (bands) of DNA, indicating how far these fragments have migrated, and hence their molecular weights.

autosome. A chromosome other than the X and Y sex chromosomes.

band. See autoradiograph.

band shift. Movement of DNA fragments in one lane of a gel at a different rate than fragments of an identical length in another lane, resulting in the same pattern “shifted” up or down relative to the comparison lane. Band shift does not necessarily occur at the same rate in all portions of the gel.

base pair (bp). Two complementary nucleotides bonded together at the matching bases (A and T or C and G) along the double helix “backbone” of the DNA molecule. The length of a DNA fragment often is measured in numbers of base pairs (1 kilobase (kb) = 1,000 bp); base-pair numbers also are used to describe the location of an allele on the DNA strand.

Bayes’ factor. A ratio that indicates how much the data change a person’s prior odds according to Bayes’ rule. In simple situations, the Bayes’ factor has the same value as the likelihood ratio.

Bayes’ rule. A formula that relates certain conditional probabilities. It can be used to describe the impact of new data on the probability that a hypothesis is true. See David H. Kaye & Hal S. Stern, *Reference Guide on Statistics and Research Methods*, in this manual.

bin, fixed. In VNTR profiling, a bin is a range of base pairs (DNA fragment lengths). When a database is divided into fixed bins, the proportion of bands within each bin is determined and the relevant proportions are used in estimating the profile frequency.

bin, floating. In VNTR profiling, a bin is a range of base pairs (DNA fragment lengths). In a floating bin method of estimating a profile frequency, the bin is centered on the base-pair length of the allele in question, and the width of the bin can be defined by the laboratory’s matching rule (e.g., $\pm 5\%$ of band size).

binning. Grouping VNTR alleles into sets of similar sizes because the alleles’ lengths are too similar to differentiate.

biogeographic ancestry testing (BGA, BGAT). Inferring the geographic region where a person’s ancestors lived from DNA variants. Clinical and anthropological research studies often use thousands of SNPs for this purpose. Much smaller panels of these ancestry-informative markers have been developed for biological trace evidence.

blind proficiency test. See proficiency test.

capillary electrophoresis. A method for separating DNA fragments (including STRs) according to their lengths. A long, narrow tube is filled with an entangled polymer or comparable sieving medium, and an electric field is applied to pull DNA fragments placed at one end of the tube through the medium. The procedure is faster and uses smaller samples than gel electrophoresis, and it can be automated.

ceiling method. A procedure for setting a minimum DNA profile frequency proposed in 1992 by a committee of the National Academy of Sciences but not recommended by a second committee in 1996. One hundred persons from each of 15 to 20 genetically homogeneous populations spanning the range of racial groups in the United States are sampled. For each allele, the higher frequency among the groups sampled (or 5%, whichever is larger) is used in calculating the profile frequency. The committees called this procedure, intended to overstate the population frequency of a profile, the ceiling principle. Compare interim ceiling principle.

centimorgan (cM). A unit of measure for the frequency of genetic recombination. One centimorgan is equal to a 1% chance that two markers on a chromosome will become separated from one another owing to a recombination event during meiosis (which occurs during the formation of egg and sperm cells).

chip. A miniaturized system for genetic analysis. One such microfluidic chip mimics capillary electrophoresis and related manipulations. DNA fragments, pulled by small voltages, move through tiny channels etched into a small block of glass, silicon, quartz, or plastic. This system is useful in analyzing STRs. Another technique places a dense array of oligonucleotide probes on a solid surface. Such hybridization microarrays are useful in identifying SNPs and in sequencing mitochondrial DNA.

chromosome. A rodlike structure composed of DNA, RNA, and proteins. Most normal human cells contain 46 chromosomes, 22 autosomes and a sex chromosome (X) inherited from the mother, and another 22 autosomes and a sex chromosome (either X or Y) inherited from the father. The genes are located along the chromosomes. See also homologous chromosomes.

coding and noncoding DNA. The sequence of base pairs in a gene that correspond to the building blocks (amino acids) of a protein is called coding DNA. (A sequence of three base pairs, called a codon, specifies a particular one of the 20 possible amino acids in the protein. The mapping of a set of three nucleotide bases to a particular amino acid is the genetic code. The cell makes the protein through intermediate steps involving coding RNA transcripts.) About 1.5% of the human genome codes for the amino-acid

sequences. Another 23.5% of the genome is classified as genetic sequence but does not encode proteins. This portion of the noncoding DNA is involved in regulating the activity of genes. It includes promoters, enhancers, and repressors. Other gene-related DNA consists of introns (that interrupt the coding sequences, called exons, in genes and that are edited out of the RNA transcript for the protein), pseudogenes (evolutionary remnants of once-functional genes), and gene fragments. The remaining extragenic DNA (about 75% of the genome) also is noncoding.

CODIS (combined DNA index system). A collection of databases on STR and other loci of DNA from convicted felons, arrestees, missing persons, and crime scenes maintained by the FBI.

complementary sequence. The sequence of nucleotides on one strand of DNA that corresponds to the sequence on the other strand. For example, if one sequence is CTGAA, the complementary bases are GACTT.

control region. See D-loop.

cytoplasm. A jelly-like material (80% water) that fills the cell.

cytosine (C). One of the four bases, or nucleotides, that make up the DNA double helix. Cytosine binds only to guanine. See nucleotide.

databank. A collection of DNA samples; a sample repository.

database. A computer-searchable collection of DNA profiles.

degradation. The breaking down of DNA by chemical or physical means.

denature, denaturation. The process of splitting, as by heating, two complementary strands of the DNA double helix into single strands in preparation for hybridization with biological probes.

deoxyribonucleic acid (DNA). The molecule that contains genetic information. DNA is composed of nucleotide building blocks, each containing a base (A, C, G, or T), a phosphate, and a sugar. These nucleotides are linked together in a double helix—two strands of DNA molecules paired up at complementary bases (A with T, C with G). See adenine, cytosine, guanine, thymine.

diploid number. See haploid number.

D-loop. A portion of the mitochondrial genome known as the control region or displacement loop that is instrumental in the regulation and initiation of mtDNA gene products. Two short “hypervariable” regions within the D-loop that do not appear to be functional are used in identity or kinship testing.

DNA polymerase. The enzyme that catalyzes the synthesis of double-stranded DNA.

DNA probe. See probe.

DNA profile. A list of the alleles at each locus. See genotype.

DNA sequence. The ordered list of base pairs in a duplex DNA molecule or of bases in a single strand.

DQ. The antigen that is the product of the DQA gene. See DQA, human leukocyte antigen.

DQA. The gene that codes for a particular class of human leukocyte antigen (HLA). This gene has been sequenced completely and can be used for forensic typing. See human leukocyte antigen.

electropherogram. The PCR products separated by capillary electrophoresis can be labeled with a dye that glows at a given wavelength in response to light shined on it. As the tagged fragments pass the light source, an electronic camera records the intensity of the fluorescence. Plotting the intensity as a function of time produces a series of peaks, with the shorter fragments producing peaks sooner. The intensity is measured in relative fluorescent units and is proportional to the number of glowing fragments passing by the detector. The graph of the intensity over time is an electropherogram.

electrophoresis. See capillary electrophoresis, gel electrophoresis.

endonuclease. An enzyme that cleaves the phosphodiester bond within a nucleotide chain.

environmental insult. Exposure of DNA to external agents such as heat, moisture, and ultraviolet radiation, or chemical or bacterial agents. Such exposure can interfere with the enzymes used in the testing process or otherwise make DNA difficult to analyze.

enzyme. A protein that catalyzes a reaction (speeds it up without being consumed).

epigenetic. Heritable changes in phenotype (appearance) or gene expression caused by mechanisms other than changes in the underlying DNA sequence. Epigenetic marks are molecules attached to DNA that can determine whether genes are active and used by the cell.

ethidium bromide. A molecule that can intercalate into DNA double helices when the helix is under torsional stress. Used to identify the presence of DNA in a sample by its fluorescence under ultraviolet light.

exon. See coding and noncoding DNA.

fallacy of the transposed conditional. See transposition fallacy.

false match. Two samples of DNA that have different profiles could be declared to match if, instead of measuring the distinct DNA in each sample, there is an error in handling or preparing samples such that the DNA from a single sample is analyzed twice. The resulting match, which does not reflect the true profiles of the DNA from each sample, is a false match. Some people use “false match”

more broadly, to include cases in which the true profiles of each sample are the same, but the samples come from different individuals. Compare true match. See also match, random match.

forensic DNA phenotyping (FDP). Inferring observable traits of individuals whose DNA is in a trace-evidence sample, or unknown deceased (missing) persons, directly from biological materials found at the scene. FDP is usually done with panels of SNPs in genes that influence external visible traits such as eye and hair color, but the term may include biogeographic ancestry testing.

forensic genetic genealogy (FGG). Another term for investigative genetic genealogy.

gel, agarose. A semisolid medium used to separate molecules by electrophoresis.

gel electrophoresis. In RFLP analysis, the process of sorting DNA fragments by size by applying an electric current to a gel. The different-sized fragments move at different rates through the gel.

gene. A set of nucleotide base pairs on a chromosome that contains the “instructions” for the product that controls some cellular function such as making an enzyme. The gene is the fundamental unit of heredity; each simple gene “codes” for a specific biological characteristic.

gene frequency. The relative frequency (proportion) of an allele in a population.

genetic drift. Random fluctuation in a population’s allele frequencies from generation to generation.

genetic genealogy. Determining family history (biological relationships between or among individuals) using DNA test results in combination with traditional genealogical methods. Direct-to-consumer genetic testing companies analyze DNA samples at hundreds of thousands of single-nucleotide polymorphisms spread across the genome. The locations and lengths of shared haploblocks are used to ascertain the degree of relatedness, and documentary and other research leads to the reconstruction of relevant parts of a family tree.

genetics. The study of the patterns, processes, and mechanisms of inheritance of biological characteristics.

genome. The complete genetic makeup of an organism, including roughly 23,000 genes and many other DNA sequences in humans. The haploid human genome comprises over three billion nucleotide base pairs.

genotype. The particular forms (alleles) of a set of genes possessed by an organism (as distinguished from phenotype, which refers to how the genotype expresses itself, as in physical appearance). In DNA analysis, the term is applied to the variations within one or all DNA regions (whether or not they constitute genes) that are analyzed.

genotype, multilocus. The alleles that an organism possesses at several sites in its genome.

genotype, single locus. The alleles that an organism possesses at a particular site in its genome.

guanine (G). One of the four bases, or nucleotides, that make up the DNA double helix. Guanine binds only to cytosine. See nucleotide.

haploblock (haplotype block). Several neighboring, tightly linked SNPs inherited together as a block. A haploblock has a higher discrimination power than any individual SNP within the block.

haploid number. Human sex cells (egg and sperm) contain 23 chromosomes each. This is the haploid number. When a sperm cell fertilizes an egg cell, the number of chromosomes doubles to 46. This is the diploid number.

haplotype. A specific combination of linked alleles at several loci on one chromosome. “Linked” means that the alleles tend to be inherited together because they are close to each other on the chromosome.

Hardy-Weinberg equilibrium. A condition in which the allele frequencies within a large, random, intrabreeding population are unrelated to patterns of mating. In this condition, the occurrence of alleles from each parent will be independent and have a joint frequency estimated by the product rule. See independence. Compare linkage disequilibrium.

heteroplasmy, heteroplasmy. The condition in which some copies of mitochondrial DNA sequences in the same individual have different base pairs or lengths.

heterozygous. Having a different allele at a given locus on each of a pair of homologous chromosomes. See allele. Compare homozygous.

homologous chromosomes. The 44 autosomes (nonsex chromosomes) in the normal human genome are in homologous pairs (one from each parent) that share an identical set of genes, but may have different alleles at the same loci.

homozygous. Having the same allele at a given locus on each of a pair of homologous chromosomes. See allele. Compare heterozygous.

human leukocyte antigen (HLA). Antigen (foreign body that stimulates an immune system response) located on the surface of most cells (excluding red blood cells and sperm cells). HLAs differ among individuals and are associated closely with transplant rejection. See DQA.

hybridization. Pairing up of complementary strands of DNA from different sources at the matching base-pair sites. For example, a primer with the sequence AGGTCT would bond with the complementary sequence TCCAGA on a DNA fragment.

identical by descent (IBD). Haplotypes that are identical and inherited from a common ancestor. IBD blocks are broken up by recombination during meiosis, so their expected length depends on the number of generations since the common ancestor at the locus. If the common ancestor lived a great many generations ago (ancient IBD), the individuals share very short tracts of genetic material. At the other extreme, in families, individuals who have IBD typically share very long tracts (greater than 10 centimorgans), and IBD is only defined with respect to documented common ancestry. Recent IBD is IBD between individuals of possibly undocumented relationship, and it results from common ancestry within approximately the past 30 generations.

INDEL. A general term that may refer to insertion or deletion of nucleotides in genomic DNA. Compare single nucleotide polymorphism.

independence. Two events are said to be independent if one is neither more nor less likely to occur when the other does.

investigative genetic genealogy (IGG). The application of genetic genealogy to generate a list of individuals whose DNA might match that of a trace-evidence or missing-person DNA sample. Also called forensic genetic genealogy (FGG) or forensic investigative genetic genealogy (FIGG).

interim ceiling principle. A procedure proposed in 1992 (and no longer used) by a committee of the National Academy of Sciences for setting a minimum DNA profile frequency. For each allele, the highest frequency (adjusted upward for sampling error) found in any major racial group (or 10%, whichever is higher) is used in product-rule calculations. Compare ceiling principle.

intron. See coding and noncoding DNA.

kilobase (kb). A measure of DNA length (1,000 bases).

kinship analysis. Evaluation of biological relatedness between individuals based on DNA. Kinship analysis is used in parentage testing (civil or criminal), disaster victim identification, missing persons identification, familial searching, genetic genealogy, and immigration control.

kinship index. A likelihood ratio used for kinship analysis. See paternity index.

likelihood ratio (LR). A measure of the support that an observation provides for one hypothesis as opposed to an alternative hypothesis. The likelihood ratio is computed by dividing the conditional probability of the observation given that one hypothesis is true by the conditional probability of the observation given the alternative hypothesis. For example, the likelihood ratio for the hypothesis that two DNA samples with the same correctly ascertained STR profile originated from the same individual (as opposed to originating from two unrelated individuals) is the reciprocal of the

random-match probability. Legal scholars have introduced the likelihood ratio as a measure of the probative value of evidence. Evidence that is 100 times more probable to be observed when one hypothesis is true as opposed to another has more probative value than evidence that is only twice as probable.

lineage marker. DNA features that are inherited exclusively from one parent. Lineage markers are found on sex chromosomes and in mitochondrial DNA.

linkage. The inheritance together of two or more genes or loci on the same chromosome.

linkage equilibrium. A condition in which the occurrence of alleles at different loci is independent.

locus. A location in the genome, that is, a position on a chromosome where a gene, marker, or other structure begins.

Markov chain Monte Carlo approximation. A computer simulation method that can be used to get an approximate solution for the integration that must be performed to apply Bayes' rule to a problem with many random variables. It is used in some computer programs that give ratios of probabilities of the data given the different sets of genotypes that might be present in a DNA sample.

massively parallel sequencing (MPS). Any of several high-throughput methods for determining nearly all or just parts of the nucleotide sequence of an individual's genome. Multiple DNA segments are analyzed simultaneously. There are several instruments with different designs to accomplish this. Also called next-generation sequencing (NGS). See reads.

match. The presence of the same allele or alleles in two samples. Two DNA profiles are declared to match when they are indistinguishable in genotype. For loci such as STRs, with discrete alleles, two samples match when they display the same set of alleles. For RFLP testing of VNTRs, two samples match when the pattern of the bands is similar and the positions of the corresponding bands at each locus fall within a preset distance. See match window, false match, true match.

match window. If two RFLP bands lie within a preset distance (called the match window) that reflects normal measurement error, the bands can be declared to match.

meiosis. A special type of cell division of germ cells that produces sperm or egg cells with only one copy of each chromosome (a haploid genome). In humans, body (somatic) cells are diploid, containing two sets of these chromosomes (one from each parent via the mother's egg cell and the father's sperm cell). Compare mitosis.

messenger RNA (mRNA). A type of single-stranded RNA that carries information from the DNA in a cell's nucleus to the cell's watery interior, where the protein-making machinery reads the mRNA sequence and translates it into the amino acid in a growing protein chain. See coding and noncoding DNA.

methylation. A chemical reaction in the body in which a small molecule called a methyl group gets added to DNA, proteins, or other molecules. The addition of methyl groups can affect how some molecules act in the body. Methylation turns gene expression on or off. See epigenetic.

microarray. See chip.

microfluidics. The technology of manufacturing microminiaturized devices containing chambers and tunnels through which fluids flow or are confined.

microhaplotype. Microhaplotype loci (microhaps) are markers less than 300 nucleotides long that display multiple allelic combinations (polymorphism).

microsatellite. Another term for an STR.

minisatellite. Another term for a VNTR.

mitochondria. A structure (organelle) within nucleated (eukaryotic) cells that is the site of the energy-producing reactions within the cell. Mitochondria contain their own DNA (often abbreviated as mtDNA), which is inherited only from mother to child.

mitosis. The process by which a cell replicates its chromosomes and then segregates them, producing two identical nuclei in preparation for cell division. Mitosis is generally followed by equal division of the cell's content into two daughter cells that have identical genomes. The copying of DNA from a human somatic (body) cell that splits to produce a daughter cell occurs in mitosis. Compare meiosis.

molecular weight. The weight in grams of 1 mole (approximately 6.02×10^{23} molecules) of a pure, molecular substance.

monomorphic. A gene or DNA characteristic that is almost always found in only one form in a population. Compare polymorphism.

multilocus probe. A probe that marks multiple sites (loci). RFLP analysis using a multilocus probe will yield an autorad showing a striped pattern of thirty or more bands. Such probes are no longer used in forensic applications.

multilocus profile. See profile.

multiplexing. Typing several loci simultaneously.

mutation. The process that produces a gene or chromosome set differing from the type already in the population; the gene or chromosome set that results from such a process.

nanogram (ng). A billionth of a gram.

next-generation sequencing (NGS). Any implementation of massively parallel sequencing. There are several generations of NGS. See reads.

nucleic acid. RNA or DNA.

nucleotide. A unit of DNA consisting of a base (A, C, G, or T) and attached to a phosphate and a sugar group; the basic building block of nucleic acids. See deoxyribonucleic acid.

nucleus. The membrane-covered portion of a eukaryotic cell containing most of the DNA and found within the cell's watery interior (the cytoplasm).

oligonucleotide. A synthetic polymer made up of fewer than 100 nucleotides; used as a primer or a probe in PCR. See primer.

paternity index. A number (technically, a likelihood ratio) that indicates the support that the paternity test results lend to the hypothesis that the alleged father is the biological father as opposed to the hypothesis that another man selected at random is the biological father. Assuming that the observed DNA genotypes (or phenotypes) correctly represent those of the mother, child, and alleged father tested, the number can be computed as the ratio of the probability of the genotypes (or phenotypes) under the first hypothesis to the probability under the second hypothesis. The paternity index is one type of kinship index; the siblingship index is another.

pH. A measure of the acidity of a solution.

phenotype. An observable trait, such as height, eye color, or blood group, resulting from a genotype and environmental factors.

picogram. A trillionth of a gram. A human cell contains about six picograms of DNA.

point mutation. See SNP.

polymarker. A commercially marketed set of PCR-based DNA tests for certain protein polymorphisms.

polymerase chain reaction (PCR). A process that mimics DNA's own replication processes to make up to millions of copies of short strands of genetic material in a few hours.

polymorphism. The presence of several forms of a gene or DNA characteristic in a population. Mutations give rise to polymorphisms. In a genome sequencing project, single nucleotide polymorphisms (SNPs) and DNA mutations are defined as DNA variants detectable in more than 1% or in less than 1% of the population, respectively. Compare allele.

population genetics. The study of the genetic composition of groups of individuals.

population structure. When a population is divided into subgroups that do not mix freely, that population is said to have structure. Significant structure can lead to allele frequencies being different in the subpopulations.

primer. An oligonucleotide that attaches to one end of a DNA fragment and provides a point for more complementary nucleotides to attach and replicate the DNA strand. See oligonucleotide.

probabilistic genotyping. Electropherograms display patterns of peaks that could result from different genotypes. Each possible genotype has its own probability of giving rise to the observed pattern (the data). In traditional binary matching, every genotype is said to match with a probability of 0 or 1, and a likelihood ratio for the single matching genotype (or single set of genotypes in a mixed profile) is assigned. Probabilistic genotyping allows the possible genotypes to give rise to the data with probabilities between 0 and 1. It uses statistical modeling to compute these probabilities. Likelihood ratios are reported for the possible genotypes (or combinations of genotypes that could be in a mixed profile).

probe. In forensics, a short segment of DNA used to detect certain alleles. The probe hybridizes, or matches up, to a specific complementary sequence. Probes allow visualization of the hybridized DNA, either by a radioactive tag (usually used for RFLP analyses) or a biochemical tag (usually used for PCR-based analyses).

product rule. When alleles occur independently at each locus (Hardy-Weinberg equilibrium) and across loci (linkage equilibrium), the proportion of the population with a given genotype is the product of the proportion of each allele at each locus, times factors of two for heterozygous loci.

proficiency test. A test administered at a laboratory to evaluate its performance. In a blind proficiency study, the laboratory personnel do not know that they are being tested.

prosecutor's fallacy. See transposition fallacy.

protein. Biologically important molecules made up of a linear string of building blocks called amino acids. The order in which these components are arranged is encoded in the DNA sequence of the gene that expresses the protein. See coding DNA.

pseudogenes. Genes that have been so disabled by mutations that they can no longer produce proteins. Some pseudogenes can still produce noncoding RNA.

quality assurance. A program conducted by a laboratory to ensure accuracy and reliability.

quality audit. A systematic and independent examination and evaluation of a laboratory's operations.

quality control. Activities used to monitor the ability of DNA typing to meet specified criteria.

random match. A match between a specific DNA profile and a second DNA profile drawn at random from the population. See also random-match probability.

random-match probability. The chance of a random match. As it is usually used in court, the random-match probability refers to the probability of a true match when the DNA being compared to the evidence DNA comes from a person drawn at random from the population. This random-true-match probability reveals the probability of a true match when the samples of DNA come from different, unrelated people.

random mating. The members of a population are said to mate randomly with respect to particular genes of DNA characteristics when the choice of mates is independent of the alleles.

rapid DNA. A fully automated process of developing a DNA profile from a sample without the need for a DNA laboratory or human intervention. A microfluidic lab-on-a-chip and associated software perform the extraction, amplification, separation, detection, and allele calling.

reads. Using next-generation “short-read” sequencing, DNA is broken into short fragments (typically 75–300 base pairs) that are amplified (copied) and then sequenced to produce “reads.” Bioinformatic techniques then arrange the reads into a continuous genomic sequence. “Third-generation sequencing” methods permit “long-read” sequencing (> 10,000 bp).

recombination. In general, any process in a diploid or partially diploid cell that generates new gene or chromosomal combinations not found in that cell or in its progenitors.

reference population. The population to which the source of a trace-evidence sample is thought to belong.

relative fluorescent unit (RFU). See electropherogram.

replication. The synthesis of new DNA from existing DNA. See polymerase chain reaction.

restriction enzyme. Protein that cuts double-stranded DNA at specific base-pair sequences (different enzymes recognize different sequences). See restriction site.

restriction fragment length polymorphism (RFLP). Variation among people in the length of a segment of DNA cut at two restriction sites.

restriction fragment length polymorphism (RFLP) analysis. Analysis of individual variations in the lengths of DNA fragments produced by digesting sample DNA with a restriction enzyme.

- restriction site.** A sequence marking the location at which a restriction enzyme cuts DNA into fragments. See restriction enzyme.
- reverse dot blot.** A detection method used to identify SNPs in which DNA probes are affixed to a membrane, and amplified DNA is passed over the probes to see if it contains the complementary sequence.
- ribonucleic acid (RNA).** A single-stranded molecule “transcribed” from DNA. “Coding” RNA acts as a template for building proteins according to the sequences in the coding DNA from which it is transcribed. Other RNA transcripts can be a sensor for detecting signals that affect gene expression or can be a switch for turning genes off or on, or they may be functionless.
- sequence-specific oligonucleotide (SSO) probe.** Also, allele-specific oligonucleotide (ASO) probe. Oligonucleotide probes used in a PCR-associated detection technique to identify the presence or absence of certain base-pair sequences identifying different alleles. The probes are visualized by an array of dots rather than by the electropherograms associated with STR analysis.
- sequencing.** Determining the order of base pairs in a segment of DNA or RNA.
- short tandem repeat (STR).** A class of repetitive DNA resulting from multiple copies of virtually identical base-pair sequences, arranged in succession at a specific locus on a chromosome. The number of repeats varies from individual to individual. STRs are shorter than VNTRs.
- single-locus probe.** A probe that only marks a specific site (locus). A single-locus probe for a VNTR will produce one or two bands in an autoradiograph. Compare multilocus probe.
- SNP (single nucleotide polymorphism).** A substitution, insertion, or deletion of a single base pair at a given point in the genome. See polymorphism. Compare indel.
- SNP chip.** See chip.
- Southern blotting.** Named for its inventor, a technique by which processed DNA fragments, separated by gel electrophoresis, are transferred onto a nylon membrane in preparation for the application of biological probes.
- thymine (T).** One of the four bases, or nucleotides, that make up the DNA double helix. Thymine binds only to adenine. See nucleotide.
- transcription.** The process by which the information in a strand of DNA is copied into a new molecule of messenger RNA (mRNA). The mRNA transcript serves as a blueprint for protein synthesis during the process of translation. See messenger RNA.
- transposition fallacy.** Also called the prosecutor’s fallacy, the transposition fallacy confuses the conditional probability of *A* given *B* with that of *B* given *A*.

Few people think that the probability that a person speaks Spanish given that he or she is a citizen of Chile equals the probability that a person is a citizen of Chile given that he or she speaks Spanish. Yet many court opinions, newspaper articles, and even some expert witnesses speak of the probability of a matching DNA genotype given that someone other than the defendant is the source of the crime-scene DNA as if it were the probability of someone else being the source given the matching profile. Transposing conditional probabilities correctly requires Bayes' rule.

true match. Two samples of DNA that have the same profile should match when tested. If there is no error in the labeling, handling, and analysis of the samples and in the reporting of the results, a match is a true match. A true match establishes that the two samples of DNA have the same profile. Unless the profile is unique, however, a true match does not conclusively prove that the two samples came from the same source. Some people use "true match" more narrowly, to mean only those matches among samples from the same source. Compare false match. See also match, random match.

variable number tandem repeat (VNTR). A class of RFLPs resulting from multiple copies of virtually identical base-pair sequences, arranged in succession at a specific locus on a chromosome. The number of repeats varies from individual to individual, thus providing a basis for individual recognition. VNTRs are longer than STRs.

whole genome sequencing. Determining the order of all the nucleotides in an individual organism's DNA. Usually accomplished with a massively parallel sequencing system.

window. See match window.

X chromosome. See chromosome.

Y chromosome. See chromosome.

References on DNA

Genetics and Genomics

- John M. Archibald, *Genomics: A Very Short Introduction* (2018).
Jocelyn E. Krebs et al., *Lewin's Genes XII* (2018).
James D. Watson et al., *Molecular Biology of the Gene* (7th ed. 2014).

Forensic Genetics and Genomics

- Jo-Anne Bright & Michael D. Coble, *Forensic DNA Profiling: A Practical Guide to Assigning Likelihood Ratios* (2020).
Forensic DNA Interpretation (John Buckleton et al. eds., 2d ed. 2016).
John M. Butler, *Fundamentals of Forensic DNA Typing* (2009); *Advanced Topics in Forensic DNA Typing: Methodology* (2011); *Advanced Topics in Forensic DNA Typing: Interpretation* (2015).
Silent Witness: Forensic DNA Analysis in Criminal Investigations and Humanitarian Disasters (Henry Ehrlich et al. eds., 2020).
Ian W. Evett & Bruce S. Weir, *Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists* (1998).
William Goodwin et al., *An Introduction to Forensic Genetics* (2d ed. 2011).
David H. Kaye, *The Double Helix and the Law of Evidence* (2010).
Erin E. Murphy, *Inside the Cell: The Dark Side of Forensic DNA* (2015).
National Research Council Committee on DNA Forensic Science: *An Update, The Evaluation of Forensic DNA Evidence* (1996).
National Research Council Committee on DNA Technology in Forensic Science, *DNA Technology in Forensic Science* (1992).
Royal Soc'y, *Forensic DNA Analysis: A Primer for Courts* (2017).

Reference Guide on Eyewitness Identification

THOMAS D. ALBRIGHT AND BRANDON L. GARRETT

Thomas D. Albright, Ph.D., is Professor and Director, Center for the Neurobiology of Vision, Conrad T. Prebys Chair in Vision Research at The Salk Institute for Biological Studies.

Brandon L. Garrett, J.D., is L. Neil Williams Professor of Law and Faculty Director, Wilson Center for Science and Justice at Duke University School of Law.

CONTENTS

Introduction, 364

The Promise and Peril of Eyewitness Testimony, 364

Definition of Eyewitness Identification, 365

A Word About Probabilities and Prediction, 365

Organization of This Reference Guide, 366

The Eyewitness as an Information-Processing “Instrument”, 367

How Is the Instrument Used?: A Taxonomy of Identification

Procedures, 367

Showups, 368

Photo Arrays, 368

Live Lineups, 369

Nonstandard Identification Procedures, 369

An Illustrative Case, 369

Predicting the Accuracy of Eyewitness Identification, 371

Manson v. Brathwaite: Accuracy Prediction by the Court, 372

Accuracy Prediction by the Scientific Community, 375

Law Enforcement, 376

The Courts, 377

Scientific Research, 377

Scientific Questions Regarding Eyewitness Identification, 378

Basic Science of Vision, Memory, and Choice, 380

How Vision Works, 380

Upper Bounds on Visual Performance, 380

Visual sensitivity, 381

Visual attention, 386

Categorical perception, 390

The Constructive Nature of Visual Experience, 392

How Memory Works, 396

Functional Processes of Memory, 396

- Memory Encoding, 396
- Memory Storage, 397
 - How memory goes missing: The “forgetting function”, 397
 - How memory goes wrong: Source memory failure and false memories, 398
 - The effect of emotion on memory storage, 399
- Memory Retrieval, 401
- Recognition Memory, 402
- How People Make Decisions, 404
- Applied Eyewitness Studies in the Laboratory, 405
 - Overview and Significance, 405
 - Performance Measures, 406
 - Discrimination, 406
 - Accuracy, 407
- Estimator Variables, 407
 - Weapon Focus, 407
 - Cross-Race Effect, 408
 - Viewing Distance, Lighting, and Exposure Duration, 409
 - Multiple Perpetrators, 412
 - Facial Recognition Ability, 413
 - Retention Interval, 413
 - Stress and Emotion, 414
 - Confidence, 416
 - Timing of identification and conditions of confidence assessment, 417
- System Variables, 419
 - Communication Between Law Enforcement Personnel and Witness, 419
 - Statements that convey prior probability of suspicion, 419
 - Communication during lineups: The importance of blinded lineup administration, 419
 - Video recording of lineup procedures, 421
 - Postidentification feedback, 421
 - Communication Between Community and Witness:
 - Social Influences on Eyewitness Performance, 422
 - The social utility of misinformation, 422
 - The long reach of modern social networks, 424
 - Evidence-Based Suspicion, 425
 - Prior Exposure and Transference Effects, 426
 - Identification Procedures, 428
 - Showups, 428
 - Confirmatory photo identification, 428
 - Lineups: Simultaneous versus sequential procedures, 429

Reference Guide on Eyewitness Identification

Other lineup procedures, 432
In-court identifications, 432
Filler Selection, 434
Corroborating Variables: Machine-Based Facial Recognition, 436
Applied Eyewitness Studies in the Field, 436
Effects of Eyewitness Testimony on Juries, 438
Sources of Expertise on Matters of Eyewitness Science, 439
The Legal Framework for Eyewitness Evidence, 441
The U.S. Supreme Court, 442
Lower Federal Court Rulings, 445
State Court Rulings, 447
State Eyewitness Evidence Statutes, 450
Police Practice Recommendations, 452
Conclusion, 453
Glossary of Terms, 454

FIGURES

1. Sensitivity to luminance contrast as a function of overall illumination, 382
2. Visual acuity declines predictably with distance from center of gaze.
Panel A: Plot of acuity as a function of eccentricity. *Panel B:* Perceptual
consequences of acuity loss, 383
3. Visual “crowding” impairs object recognition, 384
4. Viewing angle affects face recognition, 385
5. Focus of attention affects recognition of more complex objects, 387
6. Errors of perception are committed to memory because of “illusory
conjunctions” of features in a visual image, 388
7. Cross-race effect, 391
8. Perceptual uncertainty in response to visual “noise,” 393
9. Clear image will resolve uncertainty in Figure 8, 393
10. Perceptual state falls on a continuum between pure stimulus and pure
imagery, 394
11. Ebbinghaus forgetting function, 398
12. Natural variation in human face recognition ability, 403
13. Summary of empirical support for simultaneous lineup advantage, 431

Introduction

The Promise and Peril of Eyewitness Testimony

Witnessing the world and reporting what we have seen are among the most elemental of human abilities. We use these abilities to learn, to understand, and to assess what is true. We find things that we misplaced. We recognize our friends. Perhaps most importantly, we share our experiences with others. We testify to the beauty of the full moon rising, to the birth of our children, and to the terror of war.

Societies have long mined the power of this native human ability by recruiting eyewitnesses to help resolve disagreements about past events. In many cases, the outcome is of little consequence (did he swing the bat, or foul his opponent?), and we readily defer judgment to the witness. In other cases, we seek out eyewitness testimony to settle fraught conflicts over civil or criminal responsibility. A first-hand report of what happened and who was there can be extremely valuable, particularly when no physical evidence exists. But in court, a factfinder's decision rests not simply on what an eyewitness reports, but also on inference about the probability that the testimony is correct. It may not matter much if you misperceive, misremember, and misreport who was at the birthday party, but in a courtroom, being wrong about who assaulted you and hijacked the car can render a uniquely tragic form of injustice.

Eyewitness reports of criminal acts rarely involve deceit;¹ witnesses are generally well-intentioned people who offer a valued piece of information that they alone possess and believe to be true. Inaccurate testimony, should it occur, typically reflects an unwitting failure of human vision and memory, a failure that gets smoothed over, in the witness's telling, by candor and confidence. The result imperils the innocent,² leaves society at continued risk,³ and undermines public trust. However, scientific advances shed light on the accuracy of eyewitness testimony. These advances have increasingly informed legal actors, including law enforcement, lawmakers, and courts. In this guide, we summarize several decades of scientific research and legal approaches, including those that incorporate scientific insights. We also discuss more recent discoveries and the challenges and opportunities posed by new technology.

1. A noteworthy exception is “understandable” deceit, in which a witness intentionally fails to identify a suspect (e.g., a gang member) for fear of retribution.

2. Brandon L. Garrett, *Convicting the Innocent: Where Criminal Prosecutions Go Wrong* (2011), <https://doi.org/10.4159/harvard.9780674060982>.

3. Jee Park, *Eyewitness Identification and Innocence*, 64 *Loy. L. Rev.* 669, 670 (2018) (explaining that “[o]f the 158 cases [of exonerations] where the true perpetrators were identified by DNA, these actual perpetrators went on to commit 150 additional violent crimes”).

Definition of Eyewitness Identification

Eyewitness identification is defined in operational legal terms as “a naming or description by which one who has seen an event testifies from memory about the person or persons involved.”⁴ In the more analytical idioms of science, it is a type of visual object recognition task used for forensic purposes. Object recognition, in turn, is a ubiquitous function of human vision and memory, in which an observer must decide whether a currently viewed stimulus is the same as a previously viewed stimulus. People commonly perform such tasks in a variety of contexts and do so very successfully when the objects are familiar, such as one’s coffee mug or the office stapler. By contrast, the majority of eyewitness reports in criminal cases are based on events in which a witness saw an unfamiliar person or persons during a single encounter, often briefly and under highly constrained conditions. Because people are poorer at identification of unfamiliar faces, relative to familiar ones,⁵ the unfamiliar-face scenario presents a significant challenge to our criminal justice system. For purposes of this guide, eyewitness identification refers exclusively to the unfamiliar case.

A Word About Probabilities and Prediction

Probabilistic reasoning is a ubiquitous feature of courtroom decisions. Evidence is rarely certain; it varies continuously in the strength with which it is probative. In a criminal case, in order to convict, for example, a jury must infer whether eyewitness evidence is more likely under one hypothesis (the chef pulled the trigger) versus a different hypothesis (someone else pulled the trigger).

Analogously, a trial judge may apply probabilistic reasoning to make decisions about the accuracy and admissibility of evidence. The best approach to this problem is to assess conditions associated with the viewing and identification events that are known from prior empirical studies to increase or decrease accuracy. The result is a form of prediction in which the evidence is assigned (often implicitly) a probability of being true. From a continuous range of such probabilities, a judge would then apply a threshold, or decision criterion, to make a binary decision about whether to permit the evidence at trial.

Building on this framework of probabilistic inference, scientific approaches to the eyewitness accuracy problem fall into two broad categories: prospective and retrospective. The prospective approach uses novel techniques to *elicit* greater accuracy in proffered testimony. We review some of these techniques herein, but the problem faced by the courts is necessarily retrospective; the goal here is to

4. Black’s Law Dictionary 894 (11th ed. 2019).

5. Bruce Vicki et al., *Matching Identities of Familiar and Unfamiliar Faces Caught on CCTV Images*, 7 J. Exp. Psych.: Applied 207 (2001), <https://doi.org/10.1037//1076-898x.7.3.207>.

better *predict* the level of accuracy. This is a problem that the sciences of visual perception and memory are well equipped to address.

We stress these points at the outset because, as we will show, the assessment of the reliability of eyewitness evidence is not only the task of the jurors as fact-finders, but it is also central to the judge's task under the Supreme Court's 1977 *Manson v. Brathwaite* ruling on eyewitness evidence. Modern scientific research provides an empirical basis for assessing such probabilities and making predictions about the accuracy of eyewitness identifications, which is the subject of this guide.

Organization of This Reference Guide

The topic of eyewitness identification has been an intense focus of scientific research and legal policy development for decades. In this guide we have synthesized a vast amount of information from a variety of sources. We offer here a brief precis to assist with navigation of the text.

We begin with a generic description of the eyewitness and the task they are confronted with, which serves as an introduction to the problem faced by the criminal justice system. Casting the eyewitness as a biological information-processing “instrument” helps to focus on general factors that limit instrument performance under the constraints of the task.

With this orientation to the problem, we turn to accuracy prediction, which lies at the heart of any judicial decision about use of eyewitness testimony. We summarize the Supreme Court ruling in *Manson v. Brathwaite*, which recognized the significance of accuracy prediction and promoted a rule-based method for doing so. We follow this with a brief account of modern scientific approaches to prediction, which have ready application to the eyewitness problem and may improve upon the existing legal approach.

The next section provides a brief taxonomy of scientific questions that have been raised in research aimed at understanding eyewitness performance and use of eyewitness testimony by the courts. This taxonomy is followed by a dive into the science itself and the implications of this knowledge for assessment of the accuracy of eyewitness testimony.

The relevant science is rich and broad. We subdivide it into (1) basic science of vision and memory, with a focus on how these systems work and how their operational characteristics limit the accuracy of eyewitness reports; and (2) applied eyewitness science, which has experimentally manipulated variables commonly associated with the eyewitness experience, such as viewing conditions or identification procedures employed by the police, and assessed effects of these variables on identification performance. These variables—together with an understanding of how vision and memory work—provide a foundation for predicting the accuracy of eyewitness reports in criminal cases.

The final section of this guide turns to legal policy developments over the past half-century, which are intended to inform and regulate the use of eyewitness testimony. While much of this legal framework will be familiar to members of the judiciary, we show that these developments—which include court rulings, legislative actions, and changes in police practices—are increasingly rooted in the growing body of basic and applied science summarized herein.

The Eyewitness as an Information-Processing “Instrument”

From a scientific or technical perspective, it is useful to consider an eyewitness as a biological instrument that records, recovers, and reports previously encountered events—a sequence known generically as “information processing.”⁶ As for any functioning instrument, there are three pieces of knowledge needed to understand its utility for application: (1) How is it going to be used? (2) How does it work? and (3) *How well* does it work? In the following section, we briefly summarize the ways in which eyewitnesses are commonly used in criminal investigation and prosecution, and we provide an example from an actual criminal case to illustrate some of the strategies and pitfalls. In later sections on the basic sciences of vision and memory, and the applied science of eyewitness identification, we describe how the instrument works and how well it can perform the task at hand. This knowledge is critical for understanding the predictive relationship between variables associated with an eyewitness event and the accuracy of an identification.

How Is the Instrument Used?: A Taxonomy of Identification Procedures

The procedures used by law enforcement for eyewitness identification have become standardized through adherence to disseminated guidelines.⁷ There are three basic procedures in use today: (1) showups, (2) photo arrays, and (3) live lineups. In addition, there are a variety of nonstandard identification procedures that are employed occasionally by police, or by eyewitnesses themselves.

6. “Information” here refers to content acquired through the senses, retrieved from memory, or derived from comparisons thereof, which improves the observer’s ability to make predictions about the state of the world.

7. See, e.g., U.S. Dep’t of Just., *Eyewitness Identification: Procedures for Conducting Photo Arrays* (2017), <https://perma.cc/TP4F-NXCC>.

Showups

In a showup, which usually occurs at or near the crime location, police officers present a single live suspect to a witness. Showup procedures are only useful shortly after the crime, as when a person matching the witness's description is discovered in proximity to the crime scene and the witness's memory of the events is believed to be fresh. Any benefit gained by this proximity is countered, however, by the inherent suggestiveness of a showup: The procedure presents a witness with a single choice, which may cause the witness to infer that the police believe that the suspect is the perpetrator.⁸ Because of this suggestiveness, showups have been, as the U.S. Supreme Court has put it, "widely condemned."⁹

Photo Arrays

Photo arrays are the most commonly used eyewitness identification procedure. In this case, the eyewitness is presented with six facial photographs,¹⁰ each in a standard portrait view. One photo is that of a suspect; the other five photos are of "fillers"—people known to be innocent who are drawn from a preexisting database. The fillers serve to challenge recognition memory and reduce suggestiveness. Ideally, the filler faces should be chosen to match the witness's description of the perpetrator (a common police practice),¹¹ not the appearance of the suspect, because the witness's description is a direct reflection of what they have stored in memory from the crime scene.

Photo arrays are of two types: simultaneous and sequential. In a simultaneous presentation, officers display all facial photos at the same time, commonly in a two by three arrangement known as a "six-pack" photo array. In a sequential presentation, by contrast, photos are displayed one at a time.¹² Problems with suggestiveness in photo arrays may still arise if fillers are not carefully chosen. For example, a lineup that includes some fillers who do not match the witness's description of the perpetrator reduces the number of sensible choices.

8. See Nancy Steblay et al., *Eyewitness Accuracy Rates in Police Showup and Lineup Presentations: A Meta-Analytic Comparison*, 27 *Law & Hum. Behav.* 523, 523–24 (2003), <https://doi.org/10.1023/a:1025438223608>.

9. *Stovall v. Denno*, 388 U.S. 293, 302 (1967).

10. Six is the standard used in the United States today, where it is deemed sufficient to challenge recognition memory. But there is not universal agreement on this number: In the United Kingdom, for example, lineups ("identity parades") are commonly composed of nine faces.

11. Police Exec. Research Forum, *A National Survey of Eyewitness Identification Procedures in Law Enforcement Agencies* (2013), <https://perma.cc/REX2-Y29T>.

12. Nat'l Research Council, *Identifying the Culprit: Assessing Eyewitness Identification* 24 (2014) [hereinafter *Identifying the Culprit*], <https://perma.cc/4246-37D2>.

Live Lineups

Before the ready availability of standardized facial photographs, live lineups were used. In this procedure, the suspect and fillers are presented in person, either as a group (simultaneous) or one at a time (sequential). While some agencies still use live lineups, they are less common today because police may need probable cause to place a nonconsenting person in a lineup, presence of counsel may be required, and it can be practically difficult to find live fillers who fairly resemble the witness's description of the perpetrator.¹³

Nonstandard Identification Procedures

A range of additional procedures may be used in situations in which officers do not have a suspect, including showing mug books or sets of photographs or asking the eyewitness to help prepare a composite image or drawing of a culprit. Police will, on occasion, display a single photograph to a witness in an effort to confirm the identity of a person of interest. Police typically limit this “confirmatory photograph” method to situations in which the person is previously known to or acquainted with the witness. Police may sometimes take a witness to view people in the field, perhaps at a location the culprit is known to frequent. Finally, eyewitnesses may try to make identifications not arranged by police, in person or on social media, offender registries, or other online image collections. All of these situations are less controlled than a lineup procedure and potentially suggestive.

An Illustrative Case

To help illustrate the process by which eyewitness evidence emerges and its potential unfolds, and to illuminate the manifold variables that come into play when a trier of fact is evaluating the accuracy of the evidence, we begin with an example from the real world. We reference this example occasionally in the remainder of the guide in order to place scientific discussions in context.

As a juror in a lengthy criminal trial, you are listening to the prosecution recount how while walking to a friend's house, a 16-year-old student was abruptly abducted, dragged into roadside bushes, and sexually assaulted by an unknown man.¹⁴ After repeatedly punching him in the face, the victim managed to break free and run away. A passerby, in his car, saw the face of the attacker. Later, during

13. *Id.* at 25.

14. Thomas D. Albright, *Why Eyewitnesses Fail*, 114 PNAS 7758, 7758–64 (2017) [hereinafter Albright, *Why Eyewitnesses Fail*], <https://doi.org/10.1073/pnas.1706891114>.

the trial, you listen to the victim's and passerby's testimony. When asked if the victim was confident that the defendant sitting in the courtroom was the culprit, the victim replies: "Yes, I'm sure . . . I will never forget what he looks like." The passerby confirms the identification, and confirms that he has no doubts at all. As a juror, your instinct is to believe the two witnesses who observed the crime and confidently conveyed their testimony. How could you not? One saw—and courageously punched—the assailant, and the other saw the assailant just moments afterward. How could either be wrong about the attacker's identity?

The situation just described is the case of Uriah Courtney, who was sentenced to life in prison. After Courtney served eight years, however, the court granted a motion to retest DNA from the victim's clothing. The results excluded Courtney, and he was released in 2013. The DNA evidence was linked, through a hit in a DNA database, to a man who lived near the crime scene. The wrongful conviction of Uriah Courtney naturally raises questions about the accuracy of the victim's testimony and points to three general factors that often undermine criminal investigations and prosecutions based on eyewitness testimony: uncertainty, bias, and overconfidence. In simple terms, appreciation of the extent to which these three factors are in play can help improve inferences about the accuracy of eyewitness testimony.

One cause of uncertainty is the fact that the second witness and Courtney were of different races and ethnicities. It is well known that people generally have greater difficulty discriminating faces of a race different from their own (relative to same-race face discrimination),¹⁵ which manifests as disproportionately large eyewitness errors in cross-racial identifications and has significant implications for racial justice.

There were also potential sources of bias built into the identification procedure used in Uriah Courtney's case. For example, the police showed a traditional simultaneous photo array to the two eyewitnesses, who had described the culprit as white, in his mid-twenties, with brown hair and a goatee. Only two of the five fillers in this lineup had any facial hair at all—and Courtney's goatee was the most conspicuous. This lineup configuration, in conjunction with the witnesses' descriptions, naturally affected the statistics of the identification process, leading to a high prior likelihood—a bias—for picking Courtney, which both eyewitnesses did. Additional potential for bias was introduced by the fact that the officer administering the lineups knew which participant was the suspect, which allowed for inadvertent behavioral signals that could have influenced the witness's decisions.

15. Rankin W. McGugin et al., *Race-Specific Perceptual Discrimination Improvement Following Short Individuation Training with Faces*, 35 *Cognitive Sci.* 330–47 (2011), <https://doi.org/10.1111/j.1551-6709.2010.01148.x>; Hoo Heat Wong, Ian D. Stephen & David R. T. Keble, *The Own-Race Bias for Face Recognition in a Multiracial Society*, 11 *Frontiers Psych.* 208 (2020), <https://doi.org/10.3389/fpsyg.2020.00208>.

Finally, during the Courtney trial the witnesses expressed marked confidence in their identifications, which likely moved the jury to convict, despite the fact that upon the victim's first viewing of a photo lineup, she was tentative, telling police that the defendant's photo was "most similar," but that she was "not sure," and had a confidence level of 60%.¹⁶ By the time of the trial, her confidence had become inflated, while the accuracy of her testimony remained unchanged.

Uriah Courtney's wrongful conviction is but one in a lengthy and growing list that involved eyewitness misidentifications, in which accuracy of eyewitness testimony was incorrectly inferred by the court. Of the over 375 individuals in the United States who have been exonerated as of this writing based on postconviction DNA testing, approximately 70% were convicted, in part, based on eyewitness misidentifications.¹⁷ We do not know how often eyewitness identifications are conducted, but according to one estimate, they may be used in tens of thousands of cases a year.¹⁸

Predicting the Accuracy of Eyewitness Identification

As emphasized above, it is impossible for a user—law enforcement or the courts—to rationally and responsibly employ eyewitness testimony without information about the probability that the testimony is correct. But where does that information come from? There is no oracular source, and witness statements about their own accuracy are naturally suspect. As with other indeterminate systems, such as medical prognoses or weather forecasting, we must draw from other pieces of information to predict eyewitness accuracy.

In this section we briefly review principles of accuracy prediction developed by law and science communities. Before doing so, we highlight a distinction between (1) estimates of the probability that witness testimony would be accurate under a particular set of known conditions, and (2) the probability that a specific witness is accurate.¹⁹ As we will show, knowledge gained from both

16. See Uriah Courtney, Innocence Center, <https://theinnocencecenter.org/case/uriah-courtney/>.

17. See Brandon L. Garrett, Convicting the Innocent: DNA Exoneration Resource Database, available at <https://www.convictingtheinnocent.com> (providing detailed information concerning DNA exoneration cases); Garrett, *supra* note 2 (summarizing DNA exoneration data); Brandon L. Garrett, *Judging Innocence*, 108 Colum. L. Rev. 55 (2008) ("The vast majority of the exonerees (79%) were convicted based on eyewitness testimony.").

18. Identifying the Culprit, *supra* note 12 (citing Alvin G. Goldstein, June E. Chance & Gregory R. Schneller, *Frequency of Eyewitness Identification in Criminal Cases: A Survey of Prosecutors*, 27 Bull. Psychonomic Soc'y 71, 73 (1989) <https://doi.org/10.3758/bf03329902>); Garrett, *supra* note 2, at 50.

19. See David L. Faigman, John Monahan & Christopher Slobogin, *Group to Individual (G2i), Inference in Scientific Expert Testimony*, 81 U. Chi. L. Rev. 417, 417–80 (2014).

basic and applied sciences makes it possible, in principle, to offer quantitative estimates of accuracy under known conditions. By contrast, we would discourage attempts by experts to apply a numerical probability to a *specific* witness, recognizing that doing so wades into the province of the court, and there are numerous unknown variables that bear on individual cases.

Manson v. Brathwaite: Accuracy Prediction by the Court

In its landmark 1977 *Manson v. Brathwaite* ruling on the use of eyewitness evidence, the U.S. Supreme Court explicitly promoted a predictive strategy for assessing the accuracy of eyewitness reports, which was intended to provide orderly grounds for decisions about the admissibility of the evidence under the Due Process Clause. This strategy involves application of a five-factor test, in which specific conditions of witnessing and reporting are compared to five conditional standards—often termed “reliability criteria”—believed to be correlated with (and potentially causal of) the probability that a witness’s testimony is correct. We briefly review these criteria and the logic behind them here because they provide the constitutional judicial standard for admissibility of eyewitness evidence (although as discussed, federal and state rules of evidence also govern eyewitness testimony in court, as well as state statutes). For each, we suggest how growing scientific knowledge of human information processing (reviewed in greater detail below) supports a predictive strategy and may bring us closer to precisely estimating the probability that eyewitness testimony is correct.²⁰

The first *Manson* reliability criterion—“the opportunity of the witness to view the criminal at the time”—is intended to address the question of visual fidelity, or accuracy of a witness’s perceptual experience. An “opportunity to view” (commonly termed “line-of-sight visibility” by the scientific community²¹)

20. The disconnect between the *Manson* “reliability criteria” and modern sciences of vision and memory has been noted by others: Gary L. Wells & Deah S. Quinlivan, *Suggestive Eyewitness Identification Procedures and the Supreme Court’s Reliability Test in Light of Eyewitness Science: 30 Years Later*, 33 L. & Hum. Behav. 1, 1–24 (2009); Randolph N. Jonakait, *Reliable Identification: Could the Supreme Court Tell in Manson v. Brathwaite?*, 52 U. Colo. L. Rev. 511 (1981); Ruth Yacona, *Manson v. Brathwaite: The Supreme Court’s Misunderstanding of Eyewitness Identification*, 39 J. Marshall L. Rev. 539 (2006); Laura Smalarz et al., *Psychological Science on Eyewitness Identification and the U.S. Supreme Court: Reconsiderations in Light of DNA-Exonerations and the Science of Eyewitness Identification*, in *The Witness Stand and Lawrence S. Wrightsman, Jr.* (Cynthia Willis-Esqueda & Brian Bornstein eds., 2016).

21. M.L. Benedikt, *To Take Hold of Space: Isovists and Isovist Fields*, 6 Environ. & Planning B: Urb. Analytics & City Sci. 1 (1979), <https://doi.org/10.1068/b060047>.

is essential, to be sure, but the current state of vision science, much of which is focused on the very question of visual sensitivity and fidelity, is such that estimates of accuracy can now be made beyond the limiting case of what is within an observer's line of sight. In particular, observers do not simply view or not view things in their environment; there is a predictive relationship between variables that introduce uncertainty in visual experience and the accuracy of observer reports.

The second *Manson* criterion—"the witness' degree of attention"—acknowledges that attention promotes greater visual fidelity and certainty of measurement, and is thus predictive of accuracy. Modern science informs our understanding of this variable as well. Specifically, attention is a limited resource; direction of attention to one part of a visual scene routinely precludes attention to other parts. Thus, the "degree of attention" is primarily meaningful in the context of what is being attended to. As detailed under discussions of scientific research (see sections titled "Visual attention" and "Weapons focus" below), a high degree of attention to a weapon, for example, will naturally draw attentional focus away from the face, introducing uncertainty in visual experience of the face and reduction of identification accuracy. Building upon this information, it is now possible to make refined predictions of accuracy based on both the degree and the likely target of attention.

The third *Manson* criterion—"the accuracy of [the witness's] prior description of the criminal"—is based on the premise that a witness's ability to recall what the perpetrator looked like ("accuracy of his prior description") can be used to infer the probability that an identification made by the witness is correct. This is a valid premise for familiar objects, on which attention has been focused, perceptual processing is commonly "deep," and the level of detail recalled is significant.²² By contrast, modern science reveals that the correlation between recall and recognition is not strong under conditions in which unfamiliar faces are witnessed under viewing conditions constrained by time, distance, lighting, stress, and poorly focused attention.²³ All of which suggests that the predictive value of a good description may be substantially limited by the conditions of viewing.

The fourth *Manson* criterion—"the level of certainty demonstrated at the confrontation"—is rooted in the simple premise that "level of certainty," or confidence, demonstrated at the "confrontation" (at the time of the initial identification test) is a predictor of eyewitness identification accuracy. This premise matches human intuition but, as we discuss in some detail below, modern science has shown that the relationship between confidence and accuracy of recognition

22. Fergus I.M. Craik & Robert S. Lockhart, *Levels of Processing: A Framework for Memory Research*, 11(6) *J. of Verbal Learning & Verbal Behav.* 671–84 (1972).

23. Melissa Pigott & John C. Brigham, *Relationship Between Accuracy of Prior Description and Facial Recognition*, 70 *J. Applied Psych.* 547 (1985), <https://doi.org/10.1037//0021-9010.70.3.547>.

memory is nuanced. Under some well-defined conditions,²⁴ the initial confidence report may correlate with the probability that the testimony is accurate.²⁵ As we see from the Courtney case, the predictive value of this measure can grow worse with the passage of time, largely owing to outside influences (e.g., reinforcement from media and counsel) that boost confidence but naturally have no effect on accuracy, reaching a nadir at the time of the in-court identification.

The fifth *Manston* criterion—“the length of time between the crime and the confrontation”—acknowledges that time-dependent memory loss increases uncertainty of recalled information, which means that memory retention time is inversely correlated with accuracy. Because accuracy is a continuous function of time and admissibility decisions are binary, a judge must apply a time threshold for evaluation. But where does that threshold come from? Scientific studies of the time–accuracy relationship show that memory loss can be modeled as a power law function, by means of which “not only can the percentage of remaining memory strength be determined for any retention interval, but this strength estimate can be translated into an estimated probability of being correct on a fair lineup of a specified size.”²⁶ In other words, modern science can provide a quantitative, empirical foundation for the time threshold that a judge employs when evaluating evidence by the fifth *Manston* criterion.

The day-to-day judicial approach to eyewitness evidence has been largely driven over the past half-century by the predictive framework of *Manston v. Brathwaite*, at least in the federal courts. State courts, as we will discuss, have introduced more recent changes, through legislation and judicial rulings, seeking to more directly incorporate scientific insights into approaches toward eyewitness evidence. In federal courts, as in state courts, scientific insights have influenced the interpretation of the *Manston* factors and other related approaches, such as the use of expert witnesses to explain eyewitness evidence to the jury, and the use of jury instructions regarding eyewitness evidence. In the meantime, the scientific community has continued to weigh in on the important matter of inferring the probability that eyewitness testimony is correct, based primarily on new scientific discoveries in the area of human information processing and the use of new statistical tools.

24. “Pristine” lineup conditions include choosing lineup fillers such that they similarly match the witness’s description of the perpetrator, making the lineup “fair,” and imposing limits on biasing factors such as nonblinded lineup administration.

25. John T. Wixted & Gary L. Wells, *The Relationship Between Eyewitness Confidence and Identification Accuracy: A New Synthesis*, 18 Psych. Sci. Pub. Int. 10 (2017), <https://doi.org/10.1177/1529100616686966>.

26. Kenneth A. Deffenbacher et al., *Forgetting the Once-Seen Face: Estimating the Strength of an Eyewitness’s Memory Representation*, 14 J. Experimental Psych.: Applied 139, 142 (2008), <https://doi.org/10.1037/1076-898x.14.2.139>.

Accuracy Prediction by the Scientific Community

Motivated by societal concerns over wrongful convictions, which began to mount significantly with the development of efficient and effective procedures for forensic genotyping (starting approximately a decade after the Supreme Court's 1977 *Manson* ruling), a number of science–law collaborations emerged to explore how new scientific knowledge and tools might improve collection and use of eyewitness testimony.²⁷ The scientific foundations continue to evolve and sharpen, as do various recommendations for practice and reform.

In 2013, the National Research Council (NRC) convened a panel of experts to consider the use and validity of eyewitness identification for forensic purposes. The panel was composed of individuals representing a variety of relevant fields, including the scientific study of visual perception and memory, sociology, statistics, law, and policing. After reviewing evidence from many sources, the panel released a comprehensive report in the fall of 2014.²⁸

Other organizations have subsequently performed reviews. Particularly noteworthy and commendable is the 2020 scientific review paper from the American Psychology–Law Society,²⁹ which covers much of the same scientific ground as the NRC report, but concentrates almost exclusively on recommendations to law enforcement for planning, designing, and conducting eyewitness identification procedures. In the remainder of this section, we focus primarily on the NRC report and its recommendations for law enforcement, for the courts, and for the advancement of eyewitness science.

The NRC report recognized the great value that eyewitness testimony can bring to criminal proceedings and the concerted efforts of law enforcement and

27. Gary L. Wells et al., *Eyewitness Identification Procedures: Recommendations for Lineups and Photospreads*, 22 Law & Hum. Behav. 603 (1998), <https://doi.org/10.1023/a:1025750605807>; U.S. Dep't of Just., Off. of Just. Programs & Nat'l Inst. of Just., *Eyewitness Evidence: A Guide for Law Enforcement* (1999), <https://perma.cc/Y9EA-B4W3>; John W. Turtle, R.C. Lindsay & Gary L. Wells, *Best Practice Recommendations for Eyewitness Evidence Procedures: New Ideas for the Oldest Way to Solve a Case*, 1(1) Canadian J. of Police & Sec. Servs. 5 (2003); *Report of the United States Court of Appeals for the Third Circuit Task Force on Eyewitness Identifications*, 92 Temp. L. Rev. 1, 6 (2019), <https://perma.cc/Q4CH-5NLU>; Gary L. Wells et al., *Policy and Procedure Recommendations for the Collection and Preservation of Eyewitness Identification Evidence*, 44 Law & Hum. Behav. 3 (2020), <https://doi.org/10.1037/lhb0000359>; Margaret Bull Kovera et al., *Science-Based Recommendations for the Collection of Eyewitness Identification Evidence*, 58 Ct. Rev. 130 (2022).

28. Identifying the Culprit, *supra* note 12 (one author of the present guide served as co-chair of the National Research Council consensus report committee, and the other author served as a committee member).

29. Wells et al., *Policy and Procedure Recommendations* (2020), *supra* note 27.

legal professionals. At the same time, however, the NRC report identified factors that may lead to wrongful conviction and made recommendations for improvement in three areas: law enforcement procedure, use of eyewitness evidence in the courts, and scientific research. With respect to eyewitness performance, the NRC recommendations are both prospective, with an eye toward eliciting greater eyewitness accuracy through procedural reforms, and retrospective, with a focus on identifying and quantifying variables that offer predictive insight into the probability that eyewitness testimony is correct.

These recommendations have been embraced by both scientific and legal communities and have had, in a few short years, a significant positive impact on the field. The NRC report itself, which is freely available from the National Academies Press, has been downloaded more than 17,000 times and has elicited rich discussion, new scientific research, and improved legal practice. In 2017, U.S. Deputy Attorney General Sally Yates issued new Department of Justice (DOJ) guidelines for the conduct of eyewitness investigations,³⁰ which hew tightly to the recommendations and language of the NRC report. Though the charter of the independent NRC defines and limits its work to be strictly advisory—unlike the DOJ, the NRC has no enforcement powers—we briefly highlight the NRC recommendations on eyewitness identification in this section, as they offer a scholarly complement to the predictive framework established by the Supreme Court in 1977.

Law Enforcement

The five NRC recommendations for law enforcement are intended to foster “adherence to guidelines that are consistent with scientific standards for data and reporting.”³¹ These include:

1. Train all law-enforcement officers in eyewitness identification
2. Implement double-blind lineup and photo array procedures
3. Develop and use standardized witness instructions
4. Document witness confidence judgments
5. Videotape the witness identification process

These law enforcement recommendations are aimed prospectively at improving the quality of evidence elicited from eyewitnesses and presented to

30. See U.S. Dep’t of Just., *supra* note 7.

31. See *Identifying the Culprit*, *supra* note 12, at 105.

the court. Adherence to these recommendations can reduce the likelihood that uncertainty, bias, and overconfidence come into play, and make it easier to retrospectively predict the probability that an identification hypothesis is true.

The Courts

The four NRC recommendations for the courts encourage adoption of procedures intended to help courts distinguish evidence that does or does not have high predictive validity. These include:

1. Conduct pretrial judicial inquiry
2. Make juries aware of prior identifications
3. Use scientific framework expert testimony
4. Use jury instructions as an alternative means to convey information

Pretrial inquiry offers an opportunity for the judge to indulge in a careful consideration of variables (those identified by *Manson*, or others) that might serve retrospectively as good predictors of the accuracy of an identification. Use of scientific framework testimony (testimony about relevant scientific facts and principles, without explicit application to the case in question) and use of “clear and concise” jury instructions both provide similar opportunities for jurors to consider variables that are good predictors of the accuracy of an identification.³² Finally, the recommendation that juries be made aware of prior identifications is intended to establish the shortest retention interval, reveal inconsistencies, and combat the influence of confidence inflation on jury deliberations. All of which should improve the quality of evidence used retrospectively to predict the accuracy of an identification.

Scientific Research

Finally, the NRC report made several recommendations for additional scientific research, recommendations aimed at developing strategies to both prospectively elicit more accurate testimony from eyewitnesses, and—for retrospective purposes—to gain greater understanding of the causal relationships between variables associated with witnessed events and accuracy of the resulting

32. Jury instruction was an explicit recommendation of the NRC eyewitness committee. Nonetheless, as summarized below in the context of scientific studies conducted in the laboratory, some research suggests that jury instructions are not always effective at preventing wrongful convictions.

testimony. As summarized in detail in the next sections, this scientific research has progressed at a rapid pace, and particularly so since the publication of the NRC report.

Scientific Questions Regarding Eyewitness Identification

The past half-century has seen the accumulation of a vast trove of scientific knowledge bearing on the validity of eyewitness testimony. This knowledge has greatly enhanced understanding of the capabilities and limitations of this form of evidence, and has advised new policies and practices. In the foregoing sections we distilled some of this science to provide ready insights for common judicial decisions. In the following sections we take a deep dive into the scientific discoveries that informed our summary, with the conviction that this knowledge will serve as a comprehensive reference that allows for greater judicial flexibility and understanding of when, why, and with what probability eyewitnesses provide accurate testimony. First, we briefly outline the types of scientific questions that are commonly addressed. This is followed by a summary of findings from: (1) basic laboratory studies that ask how brain systems for vision and memory work; (2) applied laboratory studies that specifically focus on human performance in an eyewitness task; (3) field studies that test the effectiveness of new eyewitness tasks in real criminal justice settings; and (4) studies that explore the efficacy of communication between the eyewitness and the trier of fact.

Most scientific studies of eyewitness identification conducted over the past half-century have focused on three central questions:

1. Can we improve prediction of eyewitness identification accuracy?

When an eyewitness makes an identification, the question naturally of foremost importance for the criminal justice system is: What is the probability that the eyewitness is correct? This is the retrospective question, which focuses on variables associated with things past (e.g., conditions of witnessing and collection of evidence by law enforcement) that can be used to infer the accuracy of eyewitness reports. Nearly all of the science relevant to eyewitness evidence has addressed this question, either directly or indirectly.

2. Are there procedural improvements that can increase accuracy?

The second, prospective, question seeks strategies that the criminal justice system might use to increase the probability of accurate testimony. The scientific literature is thick with empirical comparisons of performance (e.g., accuracy) elicited by different identification procedures, such as showups versus lineups, or

simultaneous versus sequential lineup presentations. The answer may favor the use by police of one procedure over the other. Most importantly for the present context, awareness by the trier of fact of these prospective improvements may serve retrospectively to help infer the probability that eyewitness testimony is correct.

3. Can we educate jurors to be better evaluators of accuracy?

The third scientific question addresses the outsized influence of eyewitness testimony on decisions by juries. The importance of this issue can be appreciated from a scientific perspective in which the use of eyewitness evidence by the courts is viewed as a process of communication: The eyewitness is the “transmitter” of information. The jury, conversely, is the “receiver” of information. If that receiver is flawed by an inability to interpret the signals that were transmitted—either by uncritical acceptance or by lack of technical knowledge—then the fidelity of information transmitted (the accuracy of the testimony) makes little difference to the outcome. It follows that the quality of judicial decisions might benefit from plain language explanations of (1) the underlying processes of vision and memory, (2) the ways that those processes can fail, contrary to intuition, and (3) why eyewitness candor and confidence are not necessarily predictive of accuracy. Carefully worded jury instructions and scientific expert testimony have been implemented in a number of jurisdictions. While some evidence challenges the effectiveness of this approach,³³ recent experiments suggest that instructions may be more effective if they explain the reasons *why* caution is critical when evaluating eyewitness accuracy, rather than simply directing caution.³⁴ We summarize the relevant science below.

33. Amanda N. Bergold et al., *Eyewitnesses in the Courtroom: A Jury-Level Experimental Examination of the Impact of the Henderson Instructions*, 17 J. Experimental Criminology 433, 433–55 (2021), <https://doi.org/10.1007/s11292-020-09412-3>; Marlee Kind Berman, *Eyewitness Identification Jury Instructions: Do They Enhance Evidence Evaluation?*, C.U.N.Y. Acad. Works (2015); Angela M. Jones & Amanda N. Bergold, *Examining the Effectiveness of the Henderson Eyewitness Instructions*, 13 J. Forensic Psych. Prac. 1, 1–24 (2017), <https://doi.org/10.1007/s11292-016-9279-6>; Brandon L. Garrett et al., *Factoring the Role of Eyewitness Evidence in the Courtroom*, 17 J. Empirical Legal Stud. 556, 556–79 (2020), <https://doi.org/10.1111/jels.12259>; Marlee Kind Dillon et al., *Henderson Instructions: Do They Enhance Evidence Evaluation?*, 17 J. Forensic Psych. Rsch. & Practice 1 (2017), <https://doi.org/10.1080/15228932.2017.1235964>; Angela M. Jones et al., *Comparing the Effectiveness of Henderson Instructions and Expert Testimony: Which Safeguard Improves Jurors' Evaluations of Eyewitness Evidence?*, 13 J. Exp. Crim. 29, 29–52 (2017), <https://doi.org/10.1007/s11292-016-9279-6>; Cindy E. Laub, Christopher D. Kimbrough & Brian H. Bornstein, *Mock Jurors' Perceptions of Eyewitnesses Versus Earwitnesses: Do Safeguards Help?*, 34(2) Am. J. Forensic Psych. 33 (2016); Athan P. Papaïliou, David V. Yokum & Christopher T. Robertson, *The Novel New Jersey Eyewitness Instruction Induces Skepticism but Not Sensitivity*, 10(12) PLoS One e0142695 (2015), <https://doi.org/10.1371/journal.pone.0142695>.

34. Brandon L. Garrett et al., *Sensitizing Jurors to Eyewitness Confidence Using “Reason-Based” Judicial Instructions*, 12(1) J. Applied Rsch. Memory & Cognition 141 (2022), <https://doi.org/10.1037/mac0000035>.

Basic Science of Vision, Memory, and Choice

Vision, memory, and decision-making have been subjects of rigorous scientific investigation since the middle of the nineteenth century, with particularly rapid advances over the past 50 years. In simple terms, the goal has been to understand how and how well the machine works. If we're considering the purchase of a new car, violin, or patio heater, the first thing we want to know is what are the use conditions under which it performs well or fails. Unlike engineered products, however, human vision and memory do not come with performance specs and a user manual. The purpose of basic research in this area is to obtain information that can enable prediction of human performance under specific conditions, such as those associated with witnessing and reporting a crime.³⁵ In the following sections, we summarize what has been learned from this basic science.

How Vision Works

Vision is usefully understood as the process of detecting informative signals about the external world and using those signals to recognize objects, make decisions, and guide behavior.³⁶ Our focus here is on aspects of visual function that place constraints on what can be sensed and perceived. In two sections, we address (1) characteristics of vision that place upper bounds on performance, and (2) vision as a constructive process that seeks to overcome input noise and uncertainty.

Upper Bounds on Visual Performance

In this section we review some specific aspects of visual processing that place upper bounds on the quantity and quality of visual information available under defined viewing conditions. This knowledge offers a starting point for evaluating the probability that an eyewitness could have seen what they said they saw.³⁷

35. See Identifying the Culprit, *supra* note 12; Albright, *Why Eyewitnesses Fail*, *supra* note 14, at 7758–64; Marc Green et al., *Forensic Vision with Application to Highway Safety* (3rd ed. 2008).

36. Charles D. Gilbert, *The Constructive Nature of Visual Processing*, in *Principles of Neural Science* 496 (Eric R. Kandel et al. eds., 6th ed. 2021); Charles D. Gilbert, *Intermediate-Level Visual Processing and Visual Primitives*, *id.* at 545; Thomas D. Albright & Winrich A. Freiwald, *High-Level Visual Processing: From Vision to Cognition*, *id.* at 564; Thomas D. Albright, *On the Perception of Probable Things: Neural Substrates of Associative Memory, Imagery, and Perception*, 74 *Neuron* 227 (2012), <https://doi.org/10.1016/j.neuron.2012.04.001>.

37. See Identifying the Culprit, *supra* note 12; Albright, *Why Eyewitnesses Fail*, *supra* note 14; Green et al., *supra* note 35.

Visual sensitivity

At the earliest stages of visual processing there are several factors that limit the visual information available for accurate perception and object recognition. Some of these factors are inherent to the visual system and largely uncontrollable, such as scattering of light by the fluid and tissues of the eye, and can be exacerbated by common observer-specific visual deficits, such as myopia, poor contrast sensitivity, or color blindness. Other factors are associated with viewing conditions, such as viewing duration and ambient illumination, which predictably influence the quantity of information that a viewer gains from a visual scene, and thus the fidelity of experience.

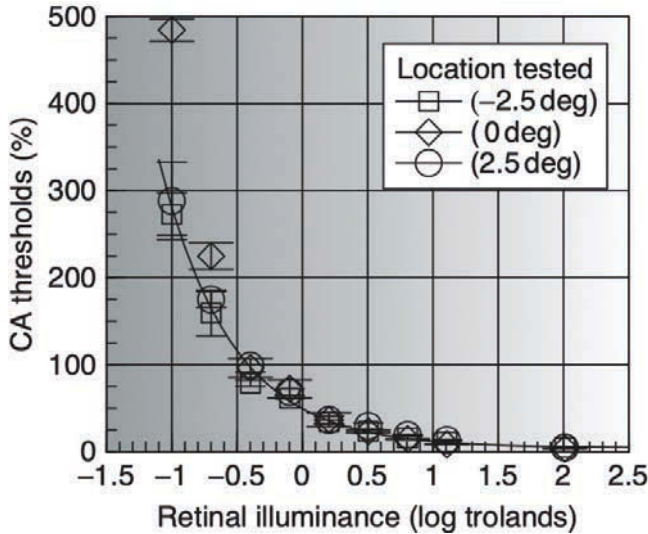
Anyone with a functioning visual system will report that their ability to resolve fine details of a viewed scene declines steeply in dim light; this is why we use flashlights, and why we cannot read a restaurant menu under candlelight. The basic science of vision puts numbers behind this experience, which can be used to predict the accuracy of eyewitness testimony.³⁸ Figure 1 shows how this effect can be quantified. The y-axis plots the amount of luminance contrast between an image feature and the background—for example, the brightness difference between a small black letter on a white page—that is required to resolve the image feature.³⁹ The x-axis plots ambient illumination. The plot shows that under very dim light (near total darkness; left side of x-axis), the image feature can be resolved only if it is extremely bright relative to background. By contrast, at high levels of illumination (daylight; right side of x-axis), feature resolution is easy. But there are intermediate levels of illumination (moonlight or candlelight; middle of x-axis), which are frequently encountered by human observers and make resolution very difficult.

The plot in Figure 1 represents one of the operating characteristics of human vision. In conjunction with measured values of ambient illumination present at a witnessed crime scene—knowable, in principle, from measurements of light sources, such as daylight, moonlight, street lights, illuminated windows, or signage—these operating characteristics can be used to predict the fidelity of experience and, in the spirit of *Manston*, assess with respectable precision whether a witness had an opportunity “to view the criminal at the time.” It is precisely for this reason that human operating characteristics—hard-won products of basic research—are such a valuable resource for accuracy predictions in the eyewitness context.

38. Percy W. Cobb, *The Influence of Illumination of The Eye on Visual Acuity: I. Introductory and Historical*, 29 *Am. J. Physiology* 76 (1911), <https://doi.org/10.1152/ajplegacy.1911.29.1.76>; Simon Shlaer, *The Relation Between Visual Acuity and Illumination*, 21 *J. Gen. Physiology* 165 (1937), <https://doi.org/10.1085/jgp.21.2.165>.

39. J.L. Barbur & A. Stockman, *Photopic, Mesopic and Scotopic Vision and Changes in Visual Performance*, in *Encyclopedia of the Eye*, vol. 3, 323 (Darlene A. Dartt et al. eds., 2010), <https://doi.org/10.1016/b978-0-12-374203-2.00233-5>.

Figure 1. Sensitivity to luminance contrast as a function of overall illumination.



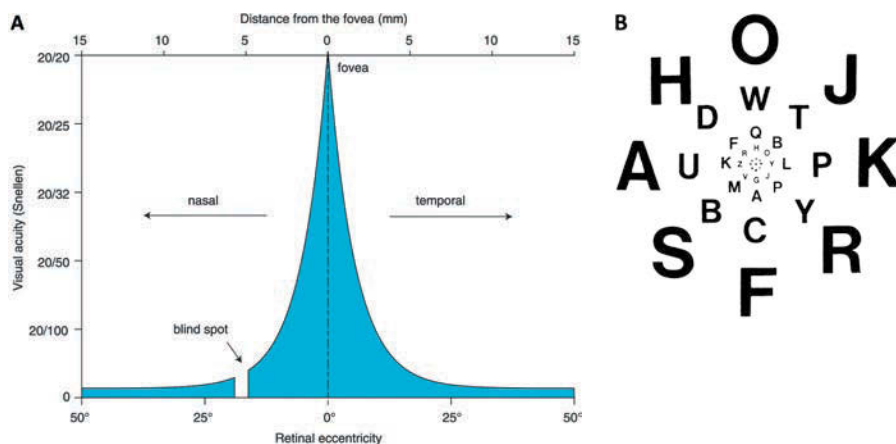
Source: Used with permission of Elsevier Science & Technology Journals, from J. L. Barbur & A. Stockman, *supra* note 39, at 323 (permission conveyed through Copyright Clearance Ctr., Inc.).

Another compelling example comes from quantitative information about visual acuity.⁴⁰ In Figure 2A, the y-axis is a measure of visual acuity. The x-axis is a measure of angular distance from the center of gaze. Normal maximum acuity is 20/20, which is achieved at the center of gaze (0 deg “retinal eccentricity” in this plot). At a viewing angle of 2 degrees from the center of gaze (about the width of a thumb at arm’s length), acuity declines by half. The perceptual consequences of this decline in acuity are illustrated in Figure 2B, where letters are scaled to be equally resolvable when directing eye gaze to the center of the image. Under time-limited, distracting, and stressful viewing conditions associated with viewing a typical crime, off-axis gaze direction—looking directly at other compelling features (such as weapon, a bleeding victim, or a second perpetrator)—can be expected to place severe constraints on the ability to detect, resolve, and memorize the face of the primary culprit, because that face image will be registered only by low-acuity regions of the sensory field.⁴¹ It follows that if one has facts about the scene, such as viewing position, distances, and angles, in conjunction

40. S.M. Anstis, *A Chart Demonstrating Variations in Acuity with Retinal Position*, 14(7) *Vision Rsch.*, 589 (1974).

41. Haus Strasburger, Ingo Rentschler & Martin Jüttner, *Peripheral Vision and Pattern Recognition: A Review*, 11 J. Vision 13 (2011), <https://doi.org/10.1167/jov.10.1167/11.5.13>.

Figure 2. Visual acuity declines predictably with distance from center of gaze. *Panel A:* Plot of acuity as a function of eccentricity. *Panel B:* Perceptual consequences of acuity loss.



Source (Panel A): Stanley Lambertus et al., *Highly Sensitive Measurements of Disease Progression in Rare Disorders: Developing and Validating a Multimodal Model of Retinal Degeneration in Stargardt Disease*, 12 PLoS One (2017), <https://doi.org/10.1371/journal.pone.0174020>. Copyright: © 2017 Lambertus et al. This is an open access article distributed under the terms of the Creative Commons Attribution License. (Panel B): Used with permission of Elsevier Science & Technology Journals, from Anstis, *supra* note 40, at 589–92 (permission conveyed through Copyright Clearance Ctr., Inc.).

with the acuity function shown above in Figure 2A, it should be possible to draw probabilistic inferences about the degree to which the witness had an opportunity “to view the criminal,” bearing (per *Manson*) on the likelihood that the witness testimony is accurate.⁴²

The ability to detect, resolve, and interpret image content is further affected in known ways by a myriad of other environmental factors, such as atmospheric conditions (rain, fog), glare, and object surface reflectance.⁴³ “Crowding,” in which objects in a visual scene are closely spaced, significantly reduces the ability

42. There is a large literature on this same topic applied to the ability of field athletes (e.g., soccer, football, tennis) to perceive critical events as a function of gaze angle, e.g., Karl Marius Aksum et al., *What Do Football Players Look At? An Eye-Tracking Analysis of the Visual Fixations of Players in 11 v. 11 Elite Football Match Play*, 11 *Frontiers Psych.* (2020), <https://doi.org/10.3389/fpsyg.2020.562995>; Francisco Belda Maruenda, *Can the Human Eye Detect an Offside Position During a Football Match?*, 329 *BMJ* 1470–72 (2004).

43. Green et al., *supra* note 35.

of observers to quickly detect and identify specific objects.⁴⁴ Figure 3 illustrates the perceptual consequence of visual crowding. In this image, letters are scaled, as in Figure 2B, to compensate for loss of visual acuity. In this case, however, the letters are “crowded” together in the central rings. As a consequence, the ability to detect and recognize specific letters is impaired. This effect suggests that a crime committed in a visually complex scene, such as a sporting event or a crowded street, will place severe and predictable limits on the ability of a witness to accurately perceive the facial features of a perpetrator.

Vision is also limited by image distortion (foreshortening) associated with angle of view. The retinal pattern generated by a face viewed directly from the front differs considerably—with changes in aspect ratio and relative placement of facial features—from that generated by a face viewed from an oblique side angle. Viewing a face from an angle above or below center (if the criminal were standing over you, or below you on the stairs) also yields retinal distortions

Figure 3. Visual “crowding” impairs object recognition.

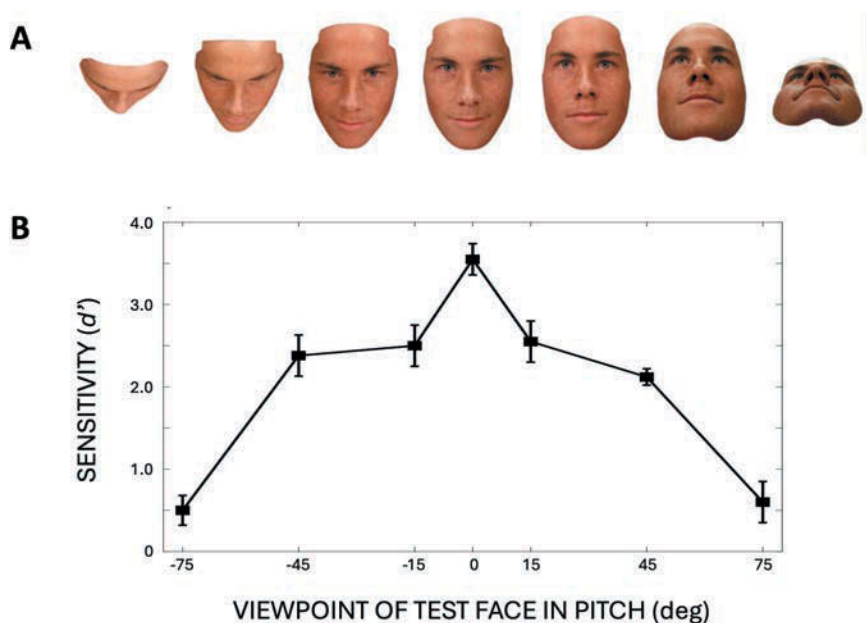


Source: Used with permission of Elsevier Science & Technology Journals, from Anstis, *supra* note 40, at 589-592 (permission conveyed through Copyright Clearance Ctr., Inc.).

44. Dennis M. Levi, *Crowding—An Essential Bottleneck For Object Recognition: A Mini-Review*, 48 *Vision Res.* 635 (2008), <https://doi.org/10.1016/j.visres.2007.12.009>; David Whitney & Dennis M. Levi, *Visual Crowding: A Fundamental Limit on Conscious Perception and Object Recognition*, 15 *Trends Cognitive Sci.* 160 (2011), <https://doi.org/10.1016/j.tics.2011.02.005>; Ryan V. Ringer et al., *Investigating Visual Crowding of Objects in Complex Real-World Scenes*, 12 *i-Perception* (2021), <https://doi.org/10.1177/2041669521994150>.

(foreshortening) of facial features, as shown in Panel A of Figure 4.⁴⁵ Panel B of Figure 4 plots the ability of human observers to discriminate (d') a frontal view (0 degrees) from the same or different face tilted up or down. Discriminability declined markedly with tilt angle. (The seven data points in Panel B correspond to the seven view angles in Panel A.) These effects are largely overcome when viewing familiar people because we have extensive prior knowledge of appearance that is independent of view angle. When viewing unfamiliar faces under time-limited conditions, however, this image distortion can severely impair subsequent recognition.⁴⁶

Figure 4. Viewing angle affects face recognition.



Note: Panel A was created by the authors using parts extracted from Figure 1 on page 3 of source. Panel B was created by the authors by modifying Figure 2, Panel A, on page 5 of source.

Source: Adapted from Simone K. Favelle & Stephen Palmisano, *The Face Inversion Effect Following Pitch and Yaw Rotations: Investigating the Boundaries of Holistic Processing*, 3 *Frontiers Psych.* 563 (2012), <https://doi.org/10.3389/fpsyg.2012.00563>. Copyright © 2012 Favelle and Palmisano. This is an open-access article distributed under the terms of the Creative Commons Attribution License.

45. Simone K. Favelle & Stephen Palmisano, *The Face Inversion Effect Following Pitch and Yaw Rotations: Investigating the Boundaries of Holistic Processing*, 3 *Frontiers Psych.* 563 (2012), <https://doi.org/10.3389/fpsyg.2012.00563>.

46. In the extreme, unconventional views of a common object, such as a bucket viewed from above, make recognition nearly impossible in the absence of context.

Visual attention

William James, the great nineteenth-century experimental psychologist, famously articulated a subjective phenomenon familiar to us all:

Everyone knows what attention is. It is the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought. Focalization, concentration, of consciousness are of its essence. It implies withdrawal from some things in order to deal effectively with others.⁴⁷

By this rule, attention, which is required for awareness, is a limited resource. Thus, only a small fraction of the information falling on the retina reaches awareness or is used by an observer for recognition, action, or storage in memory. Allocation of selective attention is an active process that can be directed by external factors—visual attributes with high salience, such as a bright light or an unfamiliar object—or internal control.⁴⁸ For example, if you are searching for your phone, you may explicitly direct your attention to the countertop where it was last seen.

Attended image content is transiently enhanced—equivalent to turning up the volume of specific sensory signals—to increase the fidelity of perceptual experience. Image content not falling within the focus of attention is processed with less fidelity.⁴⁹ In some cases, unattended content is effectively invisible: It does not reach awareness, it is not perceived, and it is not available for use in guiding decisions or actions, or for storage in memory.⁵⁰

Different pieces of visual information compete dynamically for attentional selection based on their changing physical attributes, locations in space, and relevance to the observer's needs and behavioral goals.⁵¹ An eyewitness must select what to attend to, often within a short window of time, without advance warning, in the presence of many novel objects and events, and under the confounding influences of anxiety and fear. Reductions in efficiency are common under such conditions, causing sensitivity to unattended items to be markedly reduced when there are many objects simultaneously competing for attention.⁵²

47. William James, *The Principles of Psychology* (Henry Holt ed., 1890).

48. Michael I. Posner, *Orienting of Attention*, 32 Q. J. Experimental Psych. 3 (1980), <https://doi.org/10.1080/0033558008248231>.

49. *Id.*

50. Arien Mack & Irvin Rock, *Inattention Blindness* (1998).

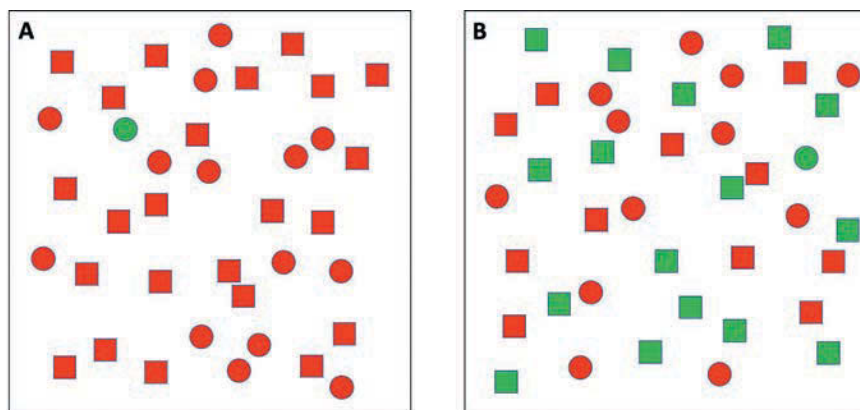
51. Harold Pashler, James C. Johnston & Eric Ruthruff, *Attention and Performance*, 52 Ann. Rev. Psych. 629 (2001), <https://doi.org/10.1146/annurev.psych.52.1.629>; Robert Desimone & John Duncan, *Neural Mechanisms of Selective Visual Attention*, 18 Ann. Rev. Neuroscience 193 (1995), <https://doi.org/10.1146/annurev.ne.18.030195.001205>.

52. James, *supra* note 47.

A simple demonstration of the complexity of attentive effects on awareness comes from the well-studied phenomenon of “visual search.”⁵³ Consider the two images in Figure 5. In panel A, the green “target” stimulus immediately “pops out” perceptually from the field of red objects and commands attention, simply because it is distinguished by a difference in color; shape varies but is irrelevant to the task. By contrast, the target stimulus in panel B is distinguished by a unique *conjunction* of color and shape (green and circular); detection of this target requires an active attentional search that proceeds serially through the objects in the display.⁵⁴ What this means, of course, is that during time-limited viewing of an unfolding crime, a witness’s attention will be automatically drawn to objects that are readily distinguished by a unique visual attribute—the woman in the red dress. Objects that are defined by unique conjunctions of features—the man on the bus in the gray sweater, surrounded by a field of men in gray coats and blue sweaters—will be harder to attend and, perhaps most importantly, harder to remember because of combinatorial complexity.

The effect of feature conjunctions on the perceptual experiences (also called percepts) that we commit to memory is illustrated by the image in Figure 6.⁵⁵ Look briefly at this image and then look away. Ask yourself whether the image contained a green letter *A*, a *B* surrounded by a diamond, or a red *D* in a circle. The red *D* in a circle is indeed present, but the *A* is blue and the *B* is surrounded

Figure 5. Focus of attention affects recognition of more complex objects.



53. Anne M. Treisman & Garry Gelade, *A Feature-Integration Theory of Attention*, 12 *Cognitive Psych.* 97 (1980), [https://doi.org/10.1016/0010-0285\(80\)90005-5](https://doi.org/10.1016/0010-0285(80)90005-5).

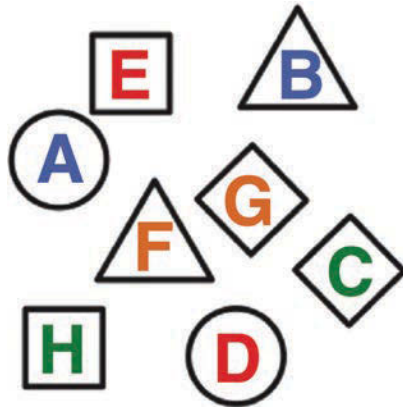
54. This search process has been described metaphorically as scanning the visual display using a “spotlight” of attention, which brings features in and out of awareness. *Id.*

55. This demonstration appears in Jeremy M. Wolfe, *Visual Search*, in *Stevens’ Handbook of Experimental Psychology and Cognitive Neuroscience* (John T. Wixted ed., 2022), <https://doi.org/10.1002/9781119170174.epcn213>.

by a triangle. Your inability to accurately report what you have seen in this demonstration reflects the visual system’s propensity to render “illusory conjunctions” of features.⁵⁶ The individual feature parameters—color, surrounding shape, and letter—are physically bound together in the image, but they are phenomenally unbound by vision, such that they easily recombine perceptually in forms that do not exist in the image. Once again, the implications for accurate eyewitness identification are significant: The person on the train platform who pulled the trigger had a mustache, brown hair, and light skin, but you may easily confuse these features with those of other people present in the scene, such that you describe him as possessing a goatee, blond hair, and light skin.

Another adverse consequence of attentional selection is that competing elements in a scene can hijack the attentional focus. The technique of misdirection—one of the essential tools of performance magic—directs attention to uninformative image content and exploits the invisibility of unattended features.⁵⁷ The well-studied phenomenon of inattention blindness is an example of misdirection,

Figure 6. Errors of perception are committed to memory because of “illusory conjunctions” of features in a visual image.



Source: Used with permission of John Wiley & Sons, Inc., from Jeremy M. Wolfe, *Visual Search*, in Stevens’ Handbook of Experimental Psychology and Cognitive Neuroscience (John T. Wixted ed., 2022), <https://doi.org/10.1002/9781119170174.epcn213>.

56. Anne Treisman & Hilary Schmidt, *Illusory Conjunctions in the Perception of Objects*, 14 *Cognitive Psych.* 107 (1982), [https://doi.org/10.1016/0010-0285\(82\)90006-8](https://doi.org/10.1016/0010-0285(82)90006-8).

57. Gustav Kuhn & Luis M. Martinez, *Misdirection—Past, Present, and the Future*, 5 *Frontiers Hum. Neuroscience* 172 (2012), <https://doi.org/10.3389/fnhum.2011.00172>; Stephen Macknik & Susana Martinez-Conde with Sandra Blakeslee, *Sleights of Mind: What the Neuroscience of Magic Reveals About Our Everyday Deceptions* (1st ed. 2010).

which occurs readily in real-world situations.⁵⁸ In this case, attention directed to one behaviorally significant property of a visual scene precludes awareness of other potentially important features.⁵⁹ As a result, an individual can be surprisingly unaware of surreptitious changes to the physical appearance of an unfamiliar person while the two are engaged in conversation.⁶⁰ This finding suggests that events that briefly divert attention have the potential to markedly impair the accuracy of eyewitness identifications.⁶¹

Attentional hijacking is particularly characteristic of stimuli that elicit strong emotional responses. Visual stimuli that trigger fear responses act as powerful external cues that command attention. While this potentiates sensitivity to those stimuli, at the expense of lost sensitivity to others, it is often the case that the attended emotional stimuli are not the ones with relevant informational content. The so-called weapon focus effect (reviewed below) is a real-world case in point for eyewitness identification, in which attention is compellingly drawn to emotionally laden stimuli, such as a gun or a knife, at the expense of acquiring greater visual information about the face of the perpetrator.⁶² Similarly, viewing a crime with multiple perpetrators (see section titled “Multiple Perpetrators” below) divides attentional resources such that the fidelity of perceptual information about each face is limited.⁶³

58. Mack & Rock, *supra* note 50; Ulric Neisser & Robert Becklen, *Selective Looking: Attending to Visually Specified Events*, 7 *Cognitive Psych.* 480 (1975), [https://doi.org/10.1016/0010-0285\(75\)90019-5](https://doi.org/10.1016/0010-0285(75)90019-5); Daniel J. Simons, *Attentional Capture and Inattention Blindness*, 4 *Trends Cognitive Scis.* 147 (2000), [https://doi.org/10.1016/s1364-6613\(00\)01455-8](https://doi.org/10.1016/s1364-6613(00)01455-8).

59. For a dramatic demonstration of this effect, produced by Daniel Simons and Christopher Chabris, see Daniel J. Simons, *Selective Attention Test*, <http://tinyurl.com/inattentional-blindness>; Daniel J. Simons & Christopher F. Chabris, *Gorillas in our Midst: Sustained Inattention Blindness for Dynamic Events*, 28 *Perception* 1059 (1999), <https://doi.org/10.1068/p281059>.

60. Daniel J. Simons & Daniel T. Levin, *Failure to Detect Changes to People During a Real-World Interaction*, 5 *Psychonomic Bull. & Rev.* 644 (1998), <https://doi.org/10.3758/bf03208840>.

61. Graham Davies & Sarah Hine, *Change Blindness and Eyewitness Testimony*, 141 *J. Psych.* 423 (2007), <https://doi.org/10.3200/jrlp.141.4.423-434>.

62. Thomas H. Kramer, Robert Buckhout & Paul Eugenio, *Weapon Focus, Arousal, and Eyewitness Memory: Attention Must be Paid*, 14 *Law & Hum. Behav.* 167 (1990), <https://doi.org/10.1007/bf01062971>; Kerri L. Pickel, Stephen J. Ross & Ronald S. Truelove, *Do Weapons Automatically Capture Attention?*, 20 *Applied Cognitive Psych.* 871 (2006), <https://doi.org/10.1002/acp.1235>; Elizabeth F. Loftus, Geoffrey R. Loftus & Jane Messo, *Some Facts About “Weapon Focus,”* 11 *Law & Hum. Behav.* 55 (1987), <https://doi.org/10.1007/bf01044839>.

63. Brian R. Clifford & Clive R. Hollin, *Effects of the Type of Incident and the Number of Perpetrators on Eyewitness Memory*, 66 *J. Applied Psych.* 364 (1981), <https://doi.org/10.1037//0021-9010.66.3.364>; Zoe J. Hobson & Rachel Wilcock, *Eyewitness Identification of Multiple Perpetrators*, 13 *Int’l J. Police Sci. & Mgmt.* 286 (2011), <https://doi.org/10.1350/ijps.2011.13.4.253>; Robert F. Lockamy et al., *One Perpetrator, Two Perpetrators: The Effect of Multiple Perpetrators on Eyewitness Identification*, 35 *Applied Cognitive Psych.* 1206 (2021), <https://doi.org/10.1002/acp.3853>; Alicia Nortje, Colin G. Tredoux & Annelies Vredeveldt, *Eyewitness Identification of Multiple Perpetrators*, 33 *S. Afr. J. Crim. Just.* 348 (2020), <https://perma.cc/CN2K-JHSY>.

Categorical perception

The precision of object recognition can easily be derailed because visual perception is categorical.⁶⁴ Although the objects of our experience vary extensively and uniquely along multiple sensory dimensions, we lump them into perceptual categories based on prior associations, many of which stem from common functions, physical properties, meanings, or emotional valence.⁶⁵ Many behavioral and cognitive processes are greatly simplified by treating all members of a category as the same, despite their physical differences. It rarely matters, for example, whether the pear we choose to eat is dappled on one side or irregular in shape, nor does the text font bear greatly on our ability to read. Evidence indicates that the structure of object memory is also categorical, suggesting that perceived objects are encoded in memory as a category type, often without specific detail, such that fine details needed for subsequent individuation are lost.⁶⁶

One of the functional corollaries of categorical perception is that observers are far better at discriminating between objects from different categories than objects from the same category. It is easier, for example, to discriminate between yellow and orange than it is to discriminate different shades of orange, even when the wavelength differences are the same.⁶⁷ Another functional corollary is that it is possible to acquire better discriminative ability—perceptual expertise—that is sharply delimited by category boundaries.⁶⁸ Through exposure and practice, we can become experts at discriminating objects within a category, such as types of sports cars, x-radiographs, and tennis shoes, that each give us little benefit when judging objects outside of that category.

Not surprisingly, a well-known phenomenon of categorical expertise is manifested as better performance on a task that involves discriminating between faces of one's own race relative to faces of a different race. This cross-race effect

64. Robert L. Goldstone, *Influences of Categorization on Perceptual Discrimination*, 123 J. Exper. Psych.: Gen. 178 (1994), <https://doi.org/10.1037//0096-3445.123.2.178>.

65. Robert R. L. Goldstone & Andrew T. Hendrickson, *Categorical Perception*, 1 Wiley Interdisc. Revs.: Cognitive Sci. 69 (2010), <https://doi.org/10.1002/wcs.26>.

66. Robert L. Goldstone, Yvonne Lippa & Richard M. Shiffrin, *Altering Object Representations Through Category Learning*, 78 Cognition 27 (2001), [https://doi.org/10.1016/s0010-0277\(00\)00099-8](https://doi.org/10.1016/s0010-0277(00)00099-8).

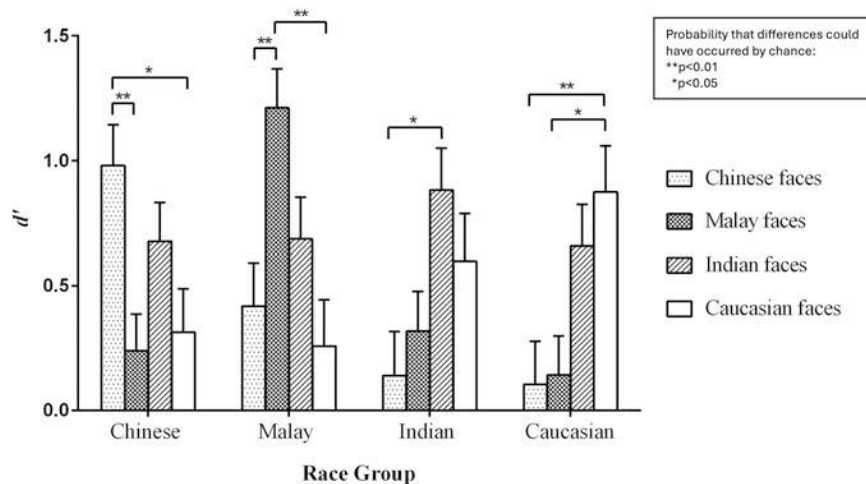
67. W. D. Wright & F. H. G. Pitt, *Hue-Discrimination in Normal Colour-Vision*, 46 Proceedings Physical Soc'y 459 (1934), <https://doi.org/10.1088/0959-5309/46/3/317>.

68. Eleanor J. Gibson, *Principles of Perceptual Learning and Development* (1969); Geoffrey R. Norman et al., *The Correlation of Feature Identification and Category Judgments in Diagnostic Radiology*, 20 Memory & Cognition 344 (1992), <https://doi.org/10.3758/bf03210919>; Rosalynn M. Peron & Gary L. Allen, *Attempts To Train Novices For Beer Flavor Discrimination: A Matter of Taste*, 115 J. Gen. Psych. 403 (1988), <https://doi.org/10.1080/00221309.1988.9710577>; Irving Biederman & Margaret M. Shiffrar, *Sexing Day-Old Chicks: A Case Study and Expert Systems Analysis of a Difficult Perceptual-Learning Task*, 13 J. Exper. Psych.: Learning, Memory & Cognition 640 (1987), <https://doi.org/10.1037//0278-7393.13.4.640>.

is pronounced and present in a variety of contexts.⁶⁹ In Figure 7, the y-axis plots a measure of face discriminability (d') for faces of four different races. The x-axis identifies the race of the human observers performing the face discrimination task. Bar texture identifies the race of the faces being discriminated. Regardless of their own race, all observers are best at discriminating between faces of their own race. (Error bars in Figure 7 represent standard errors of the mean.) The differential perceptual expertise that underlies the cross-race effect is a natural consequence of perceptual learning from exposure biases (e.g., most children grow up surrounded by family and friends of the same race),⁷⁰ though there may be some consequence of in-group versus out-group social biasing effects.⁷¹

Category-specific perceptual expertise is an operating characteristic of biological sensory systems that affords the ability to gain proficiency in specialized domains, such as medical diagnosis of skin lesions, or the ability to quickly

Figure 7. Cross-race effect.



Source: Adapted from Wong et al., *supra* note 69, at 208. Copyright © 2020 Wong, Stephen and Keeble. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY).

69. Rankin W. McGugin et al., *Race-Specific Perceptual Discrimination Improvement Following Short Individuation Training with Faces*, 35 *Cognitive Sci.* 330–47 (2011), <https://doi.org/10.1111/j.1551-6709.2010.01148.x>; Hoo Heat Wong et al., *The Own-Race Bias for Face Recognition in a Multi-racial Society*, 11 *Frontiers Psych.* 208 (2020), <https://doi.org/10.3389/fpsyg.2020.00208>.

70. McGugin et al., *supra* note 69.

71. Eric Hehman, Eric W. Mania & Samuel L. Gaertner, *Where the Division Lies: Common Ingroup Identity Moderates the Cross-Race Facial-Recognition Effect*, 46 *J. Exper. Soc. Psych.* 445 (2010), <https://doi.org/10.1016/j.jesp.2009.11.008>.

identify the gender of a baby chick.⁷² But this same category specificity has obvious potential to yield inequity of recognition decisions under conditions of uncertainty in the real world.

The Constructive Nature of Visual Experience

Any effort to assess the validity of what one says they saw must come to grips with the fact that visual perception is, by its very nature and for very good reasons, not always an accurate reproduction of the world. Vision does not operate like a camera. Rather, visual perception is a *construct* of what we predict the world to be. To understand this constructive mandate, consider the optical and biological processes that underlie visual perception: Light reflected off of objects in the world around us is refracted by the lens in the anterior portion of the eye and projected as a two-dimensional (2D) patterned image onto the back surface of the eye, known as the retina. The goal of vision is to infer, from these 2D retinal patterns, the structure and meaning of objects and events in the three-dimensional (3D) world around us. This inferential problem is constrained, however, by the fact that the projected image does not contain—indeed *cannot* contain—all of the information needed to accurately determine the real-world cause of the image.⁷³ To put it bluntly, there is an infinite set of real-world environments that can give rise to any specific retinal image.

Visual perception overcomes much uncertainty about the cause of sensation through reference to prior knowledge or experience. From these priors, or “biases,” the human brain fills in perceptual gaps with what is most likely to be out there. To illustrate this effect, consider the abstract visual pattern that appears in Figure 8. This pattern elicits a high degree of perceptual uncertainty; most people struggle to identify what it is. Figure 9 provides an additional bit of information that instills a bias. When returning to Figure 8, that bias enables the observer to improvise perceptual structure and meaning where formerly there was none.

Biases born from experience thus play a ubiquitous and essential role in perception. This concept is illustrated schematically in Figure 10. Information about the world gathered from sensory input, and information previously stored in memory, normally collaborate to yield coherent percepts. Thus, under most conditions, what we actually see falls somewhere along a continuum of mixtures of sensation and memory—a condition that facilitates behavioral interaction with the world under conditions of sensory uncertainty. The two ends of this continuum, however, are problematic states. At one end, pure stimulus is the

72. Goldstone, *supra* note 64; Biederman & Shiffrar, *supra* note 68.

73. This is a classic “inverse problem” in that there is not sufficient information in the retinal image to uniquely infer the values of the parameters that gave rise to it.

Reference Guide on Eyewitness Identification

Figure 8. Perceptual uncertainty in response to visual “noise.”



Source: Paul B. Porter, *Another Puzzle-Picture*, 67 Am. J. Psych. 550–51 (1954), <https://doi.org/10.2307/1417952>. Copyright 1954 by the Board of Trustees of the University of Illinois. Used with permission of the University of Illinois Press.

Figure 9. Clear image will resolve uncertainty in Figure 8.



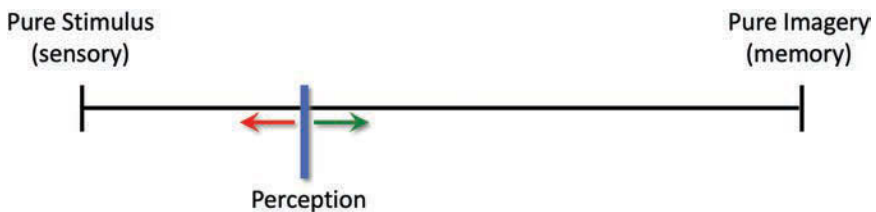
Source: Paul B. Porter, *Another Puzzle-Picture*, 67 Am. J. Psych. 550–51 (1954), <https://doi.org/10.2307/1417952>. Copyright 1954 by the Board of Trustees of the University of Illinois. Used with permission of the University of Illinois Press.

condition in which a meaningful interpretation is impossible. An observer's initial viewing of Figure 8 leads to that state. This type of perceptual failure is rare in adult visual experience, but it is debilitating and potentially dangerous when it occurs, because the observer is confronted with a stimulus for which rational decisions and actions are impossible. At the other end of the continuum, pure imagery is the hallucinatory condition in which perceptual experience is detached from all sensory input, which can be entertaining or terrifying, but not at all adaptive. There are thus selective pressures for perceptual improvisation, through integration of sensation and memory.

Though evolution has thus afforded us the ability to employ prior knowledge to perceive an unambiguous world under conditions of uncertainty, there is a well-known hazard associated with this process: Our priors are sometimes wrong. In one of the most provocative demonstrations of this phenomenon, Jerome Bruner and Leo Postman used “trick” playing cards to demonstrate a persistent biasing influence of prior knowledge on perception.⁷⁴ The trick cards were created by altering the color of a given suit—a red eight of spades, for example. Observers were shown a series of cards with brief presentations; some cards were trick and the others normal. With surprising frequency, observers failed to identify the trick cards and instead reported them as normal, thus demonstrating that strongly learned associations are capable of sharply, persistently, and unconsciously biasing visual perception. The heightened uncertainty associated with rapidly unfolding events of a crime is similarly ripe for unconscious introduction of biases held by the observer.

To make matters worse, perceptual inaccuracy born from uncertainty and bias is often associated with overconfidence—a special form of bias in which the observer implicitly rates the certainty of the experience as greater than it warrants. This phenomenon was pronounced in the “trick card” experiment cited above: Upon questioning about what they had perceived, the observers staunchly

Figure 10. Perceptual state falls on a continuum between pure stimulus and pure imagery.



74. Jerome S. Bruner & Leo Postman, *On the Perception of Incongruity: A Paradigm*, 18 J. Personality 206 (1949), <https://doi.org/10.1111/j.1467-6494.1949.tb01241.x>.

defended their mistaken identifications. Both eyewitness experiments conducted in the laboratory⁷⁵ and studies of actual criminal cases⁷⁶ demonstrate that while confidence may be low at the time of the initial lineup identification, confidence grows with reinforcement and with receipt of what appears as independent corroborating information—someone else saw something similar—which may drive the observer’s certainty in line with a larger story (see section below titled “Confidence”).⁷⁷

In summary, the sense of vision is a biological information-processing device. As with any such device, the information gained is plagued by uncertainty, bias, and overconfidence. Uncertainty is a natural consequence of the incompleteness and ambiguity of sensory information. Biases drawn from memory fill in the gaps to ensure a seamless perceptual experience. Confidence in what was perceived grows as those experiences are reinforced by other sources of information. The significance of this perspective for eyewitness testimony is profound, as it helps us to realize that the accuracy of information about the environment—the face of a criminal perpetrator, for example—gained through vision is necessarily, and often sharply, limited by ambiguity and noise. It naturally follows that having an opportunity “to view the criminal at the time” (per *Manson*) may not be sufficient for assessment of the accuracy of eyewitness testimony, because what was viewed is not necessarily what was perceived and remembered.

75. Gary L. Wells, Tamara J. Ferguson & R. C. Lindsay, *The Tractability of Eyewitness Confidence and Its Implications for Triers of Fact*, 66 J. Applied Psych. 688 (1981), <https://doi.org/10.1037//0021-9010.66.6.688>; Melissa Paiva et al., *Influence of Confidence Inflation and Explanations for Changes in Confidence on Evaluations of Eyewitness Identification Accuracy*, 16 Legal & Criminological Psych. 266 (2011), <https://doi.org/10.1348/135532510x503340>; Siegfried Ludwig Sporer et al., *Choosing, Confidence, and Accuracy: A Meta-Analysis of the Confidence-Accuracy Relation in Eyewitness Identification Studies*, 118 Psych. Bull. 315 (1995), <https://doi.org/10.1037//0033-2909.118.3.315>; Kenneth A. Deffenbacher, *Eyewitness Accuracy and Confidence: Can We Infer Anything About Their Relationship?*, 4 Law & Hum. Behav. 243 (1980), <https://doi.org/10.1007/bf01040617>; Kenneth A. Deffenbacher & William Sturgill, *Some Predictors of Eyewitness Memory Accuracy, Practical Aspects of Memory*, 4 Law & Hum. Behav. 219 (1980), <https://psycnet.apa.org/doi/10.1007/BF01040617>; Mark Tippens Reinitz et al., *Confidence–Accuracy Relations for Faces and Scenes: Roles of Features and Familiarity*, 19 Psychonomic Bull. & Rev. 1085 (2012), <https://doi.org/10.3758/s13423-012-0308-9>; Robert K. Bothwell, Kenneth A. Deffenbacher & John C. Brigham, *Correlation of Eyewitness Accuracy and Confidence: Optimality Hypothesis Revisited*, 72 J. Applied Psych. 691 (1987), <https://doi.org/10.1037//0021-9010.72.4.691>; Amy Bradfield Douglass & Eric E. Jones, *Confidence Inflation in Eyewitnesses: Seeing Is Not Believing*, 18 Legal & Criminological Psych. 152 (2020), <https://doi.org/10.1111/j.2044-8333.2011.02031.x>; Laura Smalarz & Gary L. Wells, *Do Multiple Doses of Feedback Have Cumulative Effects on Eyewitness Confidence?*, 9 J. Applied Rsch. Memory & Cognition 508 (2020), <https://doi.org/10.1037/h0101857>; Nancy K. Steblay, Gary L. Wells & Amy Bradfield Douglass, *The Eyewitness Post Identification Feedback Effect 15 Years Later: Theoretical and Policy Implications*, 20 Psych., Pub. Pol’y. & L. 1 (2014), <https://doi.org/10.1037/law0000001>; Amy Bradfield Douglass & Nancy Steblay, *Memory Distortion in Eyewitnesses: A Meta-Analysis of the Post-Identification Feedback Effect*, 20 Applied Cognitive Psych. 859 (2006), <https://doi.org/10.1002/acp.1237>.

76. Garrett, *supra* note 2.

77. Daniel Kahneman & Amos Tversky, *On the Psychology of Prediction*, 80 Psych. Rev. 237 (1973), <https://doi.org/10.1037/h0034747>.

How Memory Works

Functional Processes of Memory

Visual perceptual experiences are commonly stored as declarative memories, which are accessed and expressed as knowledge about the world. Declarative memories are of two types: Semantic memories are of facts and concepts, whereas episodic memories are of events, such as those witnessed during a crime or a walk in the park. Declarative memories are processed in three stages—encoding, storage, and retrieval—which refer to placement of items in memory, maintenance therein, and subsequent access to the stored information.

In the following sections, we briefly review memory encoding, storage, and retrieval, with emphasis on the limits of these processes as they pertain to eyewitness identification. We also discuss the phenomenon of “recognition memory,” which refers to the specific type of memory retrieval in which a stimulus (e.g., a face) is used to make an identification.

Memory Encoding

Encoding is the process through which perceived objects and events are placed into storage. This process occurs in two temporal stages, distinguished by the quantity of information stored, the duration of storage, and the susceptibility to interference.⁷⁸ Short-term or working memory is the conscious content of recent perceptual experiences, or information recently recalled from long-term storage. Short-term memory is of limited duration and capacity⁷⁹ and is labile, decaying quickly and easily disrupted by other perceptual or cognitive processes.⁸⁰

Through cellular and molecular events that play out over time, the contents of short-term memories may be encoded and consolidated into long-term memory,⁸¹ which is more enduring (albeit evolving with ongoing experience), and of greater capacity. Memories are particularly labile during this encoding process. The contents can be disrupted by interference and biased by prior

78. R. C. Atkinson & R. M. Shiffrin, *Human Memory: A Proposed System and Its Control Processes*, in 2 *The Psychology of Learning and Motivation* 89 (1968), [https://doi.org/10.1016/s0079-7421\(08\)60422-3](https://doi.org/10.1016/s0079-7421(08)60422-3); James, *supra* note 47; Alan Baddeley, *Working Memory: Looking Back and Looking Forward*, 4 *Nature Revs. Neurosci.* 829 (2003), <https://doi.org/10.1038/nrn1201>; Alan Baddeley, *Working Memory* (1986).

79. George A. Miller, *The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information*, 63 *Psych. Rev.* 81 (1956), <https://doi.org/10.1037/h0043158>.

80. John Jonides et al., *The Mind and Brain of Short-Term Memory*, 59 *Ann. Rev. Psych.* 193 (2008), <https://doi.org/10.1146/annurev.psych.59.103006.093615>.

81. Eric Kandel & Larry Squire, *Memory: From Mind to Molecules* (1999).

knowledge, expectations, or beliefs, resulting in a distorted representation of experience. Short-term memories of events that happened early in a witnessed proceeding may simply be forgotten with the passage of time or compromised by attention directed to subsequent emotional events or cognitive and behavioral demands (e.g., anxiety, fear, the need to escape). In such cases, the compromised information may never be consolidated fully into long-term storage, or that storage may contain distorted content.⁸² At the same time, the quality of encoding of stimuli that are attended is commonly enhanced by highly emotional content.⁸³

Memory Storage

Once memories have been encoded, they may be retained indefinitely in long-term storage. That stored information is not immutable, however, like information locked in a vault. On the contrary, stored memories fade with the passage of time, and they are commonly modified—often without awareness—as new knowledge and experiences are acquired. The quantitative features of these effects can provide insight into the validity of eyewitness reports.

How memory goes missing: The “forgetting function”

Much of the plasticity that plagues memory storage is the simple loss of information, which we call forgetting. The famous Ebbinghaus “forgetting function,” reproduced in Figure 11, illustrates that under the conditions used in Ebbinghaus’s nineteenth-century experiment (recall of nonsense syllables), two-thirds of the experienced content was forgotten after 24 hours.⁸⁴ The rate of forgetting

82. James L. McGaugh, *Memory—a Century of Consolidation*, 287 *Sci.* 248 (2000), <https://doi.org/10.1126/science.287.5451.248>; James L. McGaugh & Benno Roozendaal, *Role of Adrenal Stress Hormones in Forming Lasting Memories in the Brain*, 12 *Current Op. Neurobiology* 205 (2002), [https://doi.org/10.1016/s0959-4388\(02\)00306-9](https://doi.org/10.1016/s0959-4388(02)00306-9).

83. Kevin N. Ochsner, *Are Affective Events Richly Recollected or Simply Familiar? The Experience and Process of Recognizing Feelings Past*, 129 *J. Exper. Psych.: Gen.* 242 (2000), <https://doi.org/10.1037//0096-3445.129.2.242>; Deborah Talmi et al., *Immediate Memory Consequences of the Effect of Emotion on Attention to Pictures*, 15 *Learning & Memory* 172 (2008), <https://doi.org/10.1101/lm.722908>; Elizabeth A. Kensinger & Daniel L. Schacter, *Neural Processes Supporting Young and Older Adults’ Emotional Memories*, 7 *J. Cognitive Neurosci.* 1 (2008), <https://doi.org/10.1162/jocn.2008.20080>; Elizabeth A. Phelps, *Emotion and Cognition: Insights from Studies of the Human Amygdala*, 57 *Ann. Rev. Psych.* 27 (2006), <https://doi.org/10.1146/annurev.psych.56.091103.070234>.

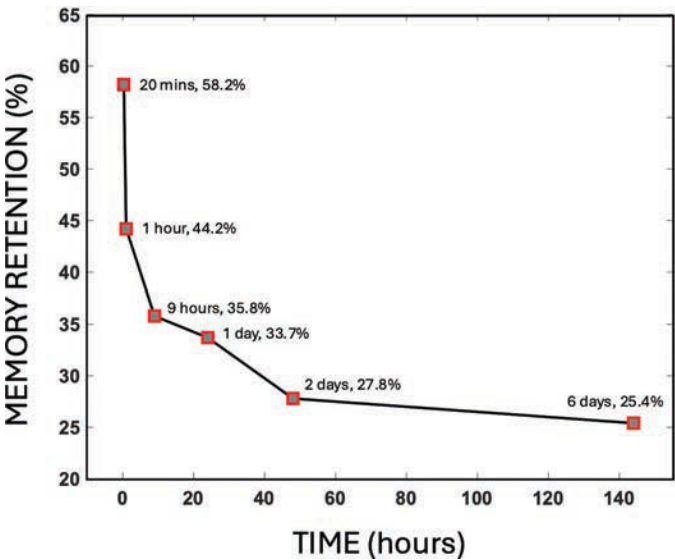
84. Hermann Ebbinghaus, *Memory: A Contribution to Experimental Psychology* (Henry A. Ruger & Clara E. Bussenius trans., 1913) (1885).

has been studied in greater detail over the past 150 years. Effects depend upon the type of memory content and the circumstances under which it was encoded, but the time-dependent decline follows a power law: $m = Lt^{-D}e^{-It}$, where m is a measure of memory retained, L is the strength of memory at time 0, t is retention time in seconds, D is decay rate, and I is interference rate.⁸⁵ The modeled value of decay captures the rapid initial decline in memory retention, which is a consequence of failure to consolidate new information into long-term memory. As consolidation proceeds with the passage of time, there is a corresponding reduction of the rate of forgetting.

How memory goes wrong: Source memory failure and false memories

Memory loss is not simply due to decay. Memory mechanisms are inherently plastic throughout life, which means that content stored for the long term is labile in the face of new information. Without awareness, we regularly reconstruct,

Figure 11. Ebbinghaus forgetting function.



85. Wayne A. Wickelgren, *Single-Trace Fragility Theory of Memory Dynamics*, 2 Memory & Cognition 775 (1974), <https://doi.org/10.3758/bf03198154>; John T. Wixted & Ebbe B. Ebbesen, *On the Form of Forgetting*, 2 Psych. Sci. 409 (1991), <https://doi.org/10.1111/j.1467-9280.1991.tb00175.x>.

update, and distort the things we believe to be true.⁸⁶ Because new content can be added and the source of that content forgotten (source memory failure, in which we effectively forget how we know things),⁸⁷ an observer may attribute updated memories to the originally witnessed events—in some cases, substantially changing what they believe they have seen. An eyewitness might learn from the police that a potential suspect has a goatee and then attribute that knowledge to the witnessed events, which could have disastrous consequences for the ability of the eyewitness to accurately report what was seen. Opportunities for this interference naturally increase with the passage of time, countering ongoing consolidation of new memories. Newly incorporated information also need not be true to fact. Research on false memories shows that it is possible to plant fabricated content in memory, which leads to recall of things that were never experienced.⁸⁸

The effect of emotion on memory storage

Memories of highly arousing emotional stimuli tend to be more enduring than memories of nonarousing stimuli.⁸⁹ Highly salient, unexpected, or arousing events—such as the *Challenger* space shuttle disaster, September 11, or episodes associated with a witnessed crime—are commonly stored strongly in memory, and their later retrieval is often associated with the subjective experience of high

86. Elizabeth F. Loftus & Geoffrey R. Loftus, *On the Permanence of Stored Information in the Human Brain*, 35 *Am. Psych.* 409 (1980), <https://doi.org/10.1037//0003-066x.35.5.409>.

87. D. Stephen Lindsay & Marcia K. Johnson, *Recognition Memory and Source Monitoring*, 29 *Bull. Psychonomic Soc'y* 203 (1991), <https://doi.org/10.3758/bf03335235>.

88. Elizabeth F. Loftus, *Planting Misinformation in the Human Mind: A 30-Year Investigation of the Malleability of Memory*, 12 *Learning & Memory* 361 (2005), <https://doi.org/10.1101/lm.94705>; Elizabeth F. Loftus & Jacqueline E. Pickrell, *The Formation of False Memories*, 25 *Psych. Annals* 720 (1995), <https://doi.org/10.3928/0048-5713-19951201-07>; Marcia K Johnson & Carol L Raye, *False Memories and Confabulation*, 2 *Trends Cognitive Scis.* 137 (1998), [https://doi.org/10.1016/s1364-6613\(98\)01152-8](https://doi.org/10.1016/s1364-6613(98)01152-8).

89. Lewis J. Kleinsmith & Stephen Kaplan, *Paired-Associate Learning as a Function of Arousal and Interpolated Interval*, 65 *J. Exper. Psych.* 190 (1963), <https://doi.org/10.1037/h0040288>; Michael W. Eysenck, *Arousal, Learning, and Memory*, 83 *Psych. Bull.* 389 (1976), <https://doi.org/10.1037//0033-2909.83.3.389>; Friderike Heuer & Daniel Reisberg, *Vivid Memories of Emotional Events: The Accuracy of Remembered Minutiae*, 18 *Memory & Cognition* 496 (1990), <https://doi.org/10.3758/bf03198482>; Tali Sharot & Elizabeth A. Phelps, *How Arousal Modulates Memory: Disentangling the Effects of Attention and Retention*, 4 *Cognitive, Affective & Behav. Neurosci.* 294 (2004), <https://doi.org/10.3758/cabn.4.3.294>; Elizabeth A. Kensinger, Rachel J. Garoff-Eaton & Daniel L. Schacter, *Memory for Specific Visual Details Can Be Enhanced by Negative Arousing Content*, 54 *J. Memory & Language* 99 (2006), <https://doi.org/10.1016/j.jml.2005.05.005>; Elizabeth A. Kensinger, *Remembering Emotional Experiences: The Contribution of Valence and Arousal*, 15 *Revs. Neurosci.* 241 (2004), <https://doi.org/10.1515/revneuro.2004.15.4.241>.

vividness and a sense of reliving.⁹⁰ The stronger encoding and storage of emotional memories results from the engagement of a specialized system of stress hormones, which is triggered by arousing content and has potentiating effects on memory consolidation and storage.⁹¹ Despite the vividness that characterizes retrieval of emotional memories, there are many indications that such memories are prone to errors.⁹² Although emotional memories are often inaccurate in detail, one important corollary of their vividness is that they are frequently held with high confidence.⁹³ This breakdown of the relationship between accuracy and confidence can undermine eyewitness accounts.⁹⁴

90. Gezinus Wolters & Joop J. Goudsmit, *Flashbulb and Event Memory of September 11, 2001: Consistency, Confidence and Age Effects*, 96 *Psych. Reps.* 605 (2005), <https://doi.org/10.2466/pr0.96.3.605-619>; Elizabeth A. Kensinger, Anne C. Krendl & Suzanne Corkin, *Memories of an Emotional and a Nonemotional Event: Effects of Aging and Delay Interval*, 32 *Exper. Aging Rsch.* 23 (2006), <https://doi.org/10.1080/01902140500325031>; Ulric Neisser & Nicole Harsch, *Phantom Flashbulbs: False Recollections of Hearing the News about Challenger, in Affect and Accuracy in Recall: Studies of "Flashbulb" Memories* 9 (Emory Symposia in Cognition, 1992), <https://doi.org/10.1017/CBO9780511664069.003>; K. S. LaBar & E. A. Phelps, *Arousal-Mediated Memory Consolidation: Role of the Medial Temporal Lobe in Humans*, 9 *Psych. Sci.* 490 (1998), <https://doi.org/10.1111/1467-9280.00090>.

91. James L. McGaugh, *Memory—A Century of Consolidation*, 287 *Sci.* 248 (2000), <https://doi.org/10.1126/science.287.5451.248>; James L. McGaugh & Benno Roozendaal, *Role of Adrenal Stress Hormones in Forming Lasting Memories in the Brain*, 12 *Current Op. Neurobiology* 205 (2002), [https://doi.org/10.1016/s0959-4388\(02\)00306-9](https://doi.org/10.1016/s0959-4388(02)00306-9).

92. Elizabeth A. Kensinger, *Remembering the Details: Effects of Emotion*, 1 *Emotion Rev.* 99 (2009), <https://doi.org/10.1177/1754073908100432>; Tali Sharot, Mauricio R. Delgado & Elizabeth A. Phelps, *How Emotion Enhances the Feeling of Remembering*, 7 *Nature Neurosci.* 1376 (2004), <https://doi.org/10.1038/nn1353>; H. Schmolck, E.A. Buffalo & L.R. Squire, *Memory Distortions Develop Over Time: Recollections of the O.J. Simpson Trial Verdict After 15 and 32 Months*, 11 *Psych. Sci.* 39 (2000), <https://doi.org/10.1111/1467-9280.00212>; Stephen R. Schmidt, *Autobiographical Memories for the September 11th Attacks: Reconstructive Errors and Emotional Impairment of Memory*, 32 *Memory & Cognition* 443 (2004), <https://doi.org/10.3758/bf03195837>; Tony W. Buchanan & Ralph Adolphs, *The Role of the Human Amygdala in Emotional Modulation of Long-Term Declarative Memory, in Emotional Cognition: From Brain to Behavior, in Advances in Consciousness Research* 9 (2002), <https://doi.org/10.1075/aicr.44.02buc>.

93. Ulrike Rimmele et al., *Emotion Enhances the Subjective Feeling of Remembering, Despite Lower Accuracy for Contextual Details*, 11 *Emotion* 553 (2011), <https://doi.org/10.1037/a0024246>; Kensinger, *supra* note 92; Neisser & Harsch, *supra* note 90; Elizabeth A. Phelps & Tali Sharot, *How (and Why) Emotion Enhances the Subjective Sense of Recollection*, 17 *Current Directions Psych. Sci.* 147 (2008) <https://doi.org/10.1111/j.1467-8721.2008.00565.x>.

94. Kate A. Houston et al., *The Emotional Eyewitness: The Effects of Emotion on Specific Aspects of Eyewitness Recall and Recognition Performance*, 13 *Emotion* 118 (2013), <https://doi.org/10.1037/a0029220>; Robin S. Edelstein et al., *Emotion and Eyewitness Memory, in Memory and Emotion* 308 (D. Reisberg & P. Hertel eds., 2004), <https://doi.org/10.1093/acprof:oso/9780195158564.003.0010>; Sven-Åke Christianson, *Emotional Stress and Eyewitness Memory: A Critical Review*, 112 *Psych. Bull.* 284 (1992), <https://doi.org/10.1037//0033-2909.112.2.284>.

Memory Retrieval

Memory retrieval refers to the process by which stored information is accessed and brought into consciousness, where it can be used to make decisions and guide actions. Retrieval of long-term declarative memories is commonly triggered through association with an external stimulus—the sound of the coffee grinder is a powerful retrieval cue for visual and olfactory memories of the morning espresso. A corollary of this association-based phenomenon is that memory retrieval is often context dependent; a memory may be more readily retrieved if the observer is in physical surroundings that are the same as or similar to those in which the original experiences took place,⁹⁵ because the surroundings provide additional cues to trigger memory retrieval.

Memory retrieval is affected by various sources of noise. Similarities of meaning or appearance between retrieval cues and items in memory can easily lead to retrieval of the wrong item, producing disruptive cognitive experiences—the sound of thunder may terrorize a veteran with retrieved memories of the battlefield—or false reports of what had been stored.⁹⁶ This is particularly a problem given the categorical nature of memory.⁹⁷ A related form of noise consists of intrusion errors, in which information known to be commonly associated with events of a general type becomes incorporated into the retrieved content of a specific memory. For example, because guns are often associated with robbery, an observer may readily and unwittingly

95. D. R. Godden & A. D. Baddeley, *Context Dependent Memory in Two Natural Environments: On Land and Underwater*, 66 *Brit. J. Psych.* 325 (1975), <https://doi.org/10.1111/j.2044-8295.1975.tb01468.x>; Stephen M. Smith & Edward Vela, *Environmental Context-Dependent Eyewitness Recognition*, 6 *Applied Cognitive Psych.* 125 (1992), <https://doi.org/10.1002/acp.2350060204>; Stephen M. Smith & Edward Vela, *Environmental Context-Dependent Memory: A Review and Meta-Analysis*, 8 *Psychonomic Bull. Rev.* 203 (2001), <https://doi.org/10.3758/bf03196157>; Endel Tulving & Donald M. Thomson, *Encoding Specificity and Retrieval Processes in Episodic Memory*, 80 *Psych. Rev.* 352 (1973), <https://doi.org/10.1037/h0020071>.

96. John R. Anderson, *A Spreading Activation Theory of Memory*, 22 *J. Verbal Learning & Verbal Behav.* 261 (1983), [https://doi.org/10.1016/s0022-5371\(83\)90201-3](https://doi.org/10.1016/s0022-5371(83)90201-3); Allan M. Collins & Elizabeth F. Loftus, *A Spreading-Activation Theory of Semantic Processing*, 82 *Psych. Rev.* 407 (1975), <https://doi.org/10.1037//0033-295x.82.6.407>; Henry L. Roediger III, David A. Balota & Jason M. Watson, *Spreading Activation and Arousal of False Memories*, in *The Nature of Remembering: Essays in Honor of Robert G. Crowder* 95 (2001), <https://doi.org/10.1037/10394-006>; C. J. Brainerd & V. F. Reyna, *The Science of False Memory* (2005), <https://doi.org/10.1093/acprof:oso/9780195154054.003.0002>.

97. Endel Tulving, *Episodic and Semantic Memory: Where Should We Go From Here?*, 9 *Behav. & Brain Scis.* 573 (1986), <https://doi.org/10.1017/s0140525x00047257>; Martha J. Farah & James L. McClelland, *A Computational Model of Semantic Memory Impairment: Modality Specificity and Emergent Category Specificity*, 120 *J. Exper. Psych.: Gen.* 339 (1991), <https://doi.org/10.1037//0096-3445.120.4.339>; Glyn W. Humphreys & Emer M. E. Forde, *Hierarchies, Similarity, and Interactivity in Object Recognition: "Category-Specific" Neuropsychological Deficits*, 24 *Behav. & Brain Scis.* 453 (2001), <https://doi.org/10.1017/s0140525x01004150>.

incorporate a gun into the retrieved version of his or her memory of a witnessed robbery.

As noted above, memory retrieval is sometimes facilitated by being in the physical surroundings where the memorized events were experienced. A related phenomenon is state-dependent memory, in which retrieval accuracy is best if the individual's cognitive state at the time of retrieval matches the cognitive state at the time of encoding.⁹⁸ When memories have an emotional component, retrieval may be best when the individual is induced to a corresponding emotional state,⁹⁹ which is accomplished by verbally or physically placing the person in the same context.¹⁰⁰

Recognition Memory

Recognition memory is the specific type of declarative memory retrieval in which an immediately present sensory stimulus elicits retrieval of the same stimulus stored following a prior encounter. The subjective determination of “sameness” is the key here, for recognition memory differs from other forms of retrieval (such as recalling a phone number or the recipe for lasagna) in that a comparison must be made between the retrieved memory and the sensory stimulus (the “cue”) that elicited retrieval. This comparison process ultimately yields a categorical decision about whether the cue and the retrieved memory of a sensory stimulus are sufficiently similar (same source) or not (different source).

Recognition memory for faces differs greatly between familiar and unfamiliar faces. Because we often identify familiar faces with ease, most people tend to think they are generally very good at face recognition. However, the ability to recognize unfamiliar faces differs widely across individuals. At one extreme are those people, referred to as “super recognizers,” who hardly forget a face. At the other end of the spectrum are “face-blind people,” who have great difficulty

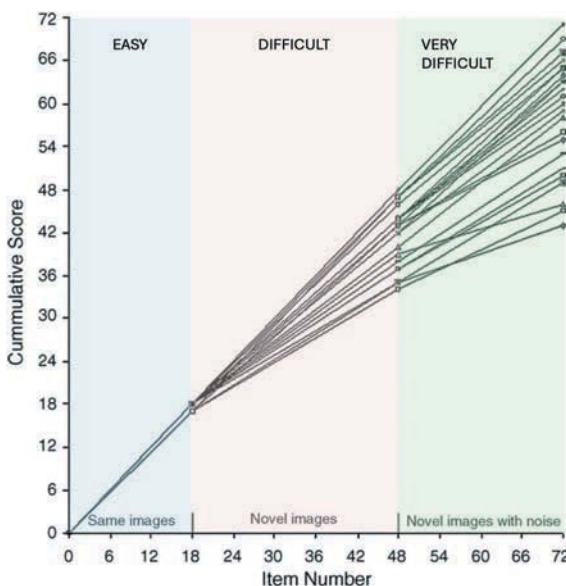
98. Donald W. Goodwin et al., *Alcohol and Recall: State-Dependent Effects in Man*, 163 *Sci.* 1358 (1969), <https://doi.org/10.1126/science.163.3873.1358>; Endel Tulving & Donald M. Thomson, *Encoding Specificity and Retrieval Processes in Episodic Memory*, 80 *Psych. Rev.* 352 (1973), <https://doi.org/10.1037/h0020071>; Edward Girden & Elmer Culler, *Conditioned Responses in Curarized Striate Muscle in Dogs*, 23 *J. Compar. Psych.* 261 (1937), <https://doi.org/10.1037/h0058634>; Donald A. Overton, *State Dependent or “Dissociated” Learning Produced with Pentobarbital*, 57 *J. Compar. & Physiological Psych.* 3 (1964), <https://doi.org/10.1037/h0048023>.

99. Pamela M. Kenealy, *Mood State-Dependent Retrieval: The Effects of Induced Mood on Memory Reconsidered*, 50 *Q. J. Exper. Psych.* 290 (1997), <https://doi.org/10.1080/713755711>; Penelope A. Lewis & Hugo D. Critchley, *Mood-Dependent Memory*, 7 *Trends Cognitive Scis.* 431 (2003), <https://doi.org/10.1016/j.tics.2003.08.005>; G. H. Bower, *Mood and Memory*, 36 *Am. Psych.* 129 (1981), <https://psycnet.apa.org/doi/10.1037/0003-066X.36.2.129>; Craik & Lockhart, *supra* note 22; Kensinger, *supra* note 92; Kenneth A. Leight & Henry C. Ellis, *Emotional Mood States, Strategies, and State-Dependency in Memory*, 20 *J. Verbal Learning & Verbal Behav.* 251 (1981), [https://doi.org/10.1016/s0022-5371\(81\)90406-0](https://doi.org/10.1016/s0022-5371(81)90406-0).

100. Smith & Vela, *A Review and Meta-Analysis*, *supra* note 95.

recognizing even highly familiar faces. Evidence indicates that the ability to recognize an unfamiliar face varies continuously between these extremes across the human population.¹⁰¹ Figure 12 shows the results of an experiment designed to demonstrate this variation. Observers performed a recognition memory task that involved matching faces of three difficulty levels, identified on the x-axis as “same” (easy), “novel” (difficult), and “novel with noise” (very difficult). The y-axis plots recognition performance (“cumulative score”), where perfect recognition corresponds to the case in which $y = x$. Each curve contains data from one of 25 subjects. For the easy category, performance was nearly equivalent across subjects and nearly perfect. For difficult faces, performance was more variable, with some subjects performing nearly perfectly and other subjects performing approximately 65% correct. For very difficult faces, performance was highly variable across subjects, with the best performers at nearly 100% correct and the

Figure 12. Natural variation in human face recognition ability.



Source: Used with permission of Elsevier Science & Technology Journals, adapted from Brad Duchaine & Ken Nakayama, *The Cambridge Face Memory Test: Results for Neurologically Intact Individuals and an Investigation of Its Validity Using Inverted Face Stimuli and Prosopagnosic Participants*, 44 *Neuropsychologia*, 576–85 (2006), <https://doi.org/10.1016/j.neuropsychologia.2005.07.001> (permission conveyed through Copyright Clearance Ctr, Inc.).

101. Brad Duchaine & Ken Nakayama, *The Cambridge Face Memory Test: Results for Neurologically Intact Individuals and an Investigation of Its Validity Using Inverted Face Stimuli and Prosopagnosic*.

worst at less than 60% correct.¹⁰² New efforts to assess where a given witness lies along this spectrum, summarized in the section below titled “Applied Eyewitness Studies in the Field,” may prove diagnostic for interpreting the validity of identification testimony.¹⁰³

How People Make Decisions

An eyewitness observer is manifestly a biological instrument for information measurement, storage, and classification. The foregoing discussions of vision and memory highlight functional capacities and operational limits of brain systems that serve the measurement and storage functions. The information measured and stored by an eyewitness is ultimately used to make a decision about identity, which is typically rendered as a binary choice (“He’s the one,” or “He’s not the one”) applied to each face in a lineup. The accuracy of that classification is an index of eyewitness performance. An understanding of the processes that underlie human decisions sheds light on where information processing works accurately or may go wrong.

Many human decisions involve comparisons of two items that have been measured by the senses. These items are often experienced as percepts derived from immediately present sensory stimuli: Which chair is bigger? Which light is brighter? Which suitcase is heavier? In some cases, the items being compared come from different informational sources, one from an immediate sensory stimulus and the other from memory of a stimulus previously experienced: Does it look like an oak tree? Does it taste like chicken? In both sensory-sensory (one sensory stimulus versus another simultaneously present sensory stimulus) and sensory-memory (a present sensory stimulus versus a remembered stimulus) comparisons, the results fall on a continuous cognitive scale of similarity. In all of these examples, however, the expected decision is not continuous; it is categorical, for example, “This one is brighter,” or “No, it does not taste like chicken.”

Human decisions thus depend upon two quantifiable elements. The first of these is the “decision variable,” which, for the examples given, is the subjective measure of similarity. An observer’s report of recognition is also influenced by the level of similarity that the observer finds acceptable. This level, known as the “decision criterion,” is thus the second element that underlies human categorical decisions. Under some conditions, the sensory and memory measures being compared are so different that the threshold level for similarity is easy to assign.

102. *Id.*

103. See Jesse H. Grabman et al., *Predicting High Confidence Errors in Eyewitness Memory: The Role of Face Recognition Ability, Decision-Time, and Justifications*, 8 J. Applied Rsch. Memory & Cognition 233, 241 (2019), <https://doi.org/10.1037/h0101835.suppl>.

There are, however, many similarity outcomes that are plausible under two different ground truth conditions (same versus different source), which make an observer's placement of the criterion for a categorical decision particularly problematic.

The decision criterion we actually apply under difficult conditions depends, in part, upon how we prioritize the outcomes. The simple desire to be informative to law enforcement, for example, may lead an eyewitness to lower their criterion for an identification decision. Estimating (or controlling) the observer's decision criterion is thus a critical step in efforts to judge the validity of an identification.

Applied Eyewitness Studies in the Laboratory

Overview and Significance

Coincident with Supreme Court rulings on eyewitness identification,¹⁰⁴ the 1970s saw a rapid growth of research on the validity of eyewitness testimony. In a field-defining review published in 1978,¹⁰⁵ Gary Wells developed a taxonomy of variables that would be expected to influence eyewitness performance, which unleashed a tide of scientific research on the effects of those variables in controlled laboratory studies designed to simulate real-world conditions.

The variables of interest, which reflect environmental and human factors, are particularly important because they are frequently at issue in criminal cases, and their known states may provide insight into the validity of eyewitness testimony. These variables fall into two classes, which Wells termed “estimator” and “system” variables.¹⁰⁶ Estimator variables characterize viewing conditions and the perceptual and cognitive states of the witness at the time of the crime. System variables, by contrast, influence identification accuracy after the crime has occurred and are variables that can be controlled by police.

The most valuable feature of applied laboratory studies of the impact of estimator and system variables is that “ground truth” is known. We thus know whether a given eyewitness made a correct or incorrect identification decision, which provides information about the predictive value of conditions associated with witnessing events (estimator variables) and identification events (system variables). The standard experimental paradigm involves exposing human

104. *Neil v. Biggers*, 409 U.S. 188 (1972); *Manson v. Brathwaite*, 432 U.S. 98, 114 (1977).

105. G. L. Wells, *Applied Eyewitness-Testimony Research: System Variables and Estimator Variables*, 36 *J. Personality & Soc. Psych.* 1546 (1978) <https://doi.org/10.1037/0022-3514.36.12.1546>.

106. *Id.*

subjects to mock crimes, in the form of staged events or videos. Subjects are then presented with a lineup and asked to report whether it contains a person they recognize from the witnessed crime. One can measure their accuracy, or whether they made a correct or incorrect choice, as well as other useful gauges of performance, such as confidence in their decision, and how quickly they arrive at that decision.

Although findings from many applied eyewitness studies have reinforced what basic research long ago revealed about the operating characteristics of vision and memory, particularly studies of the effects of viewing conditions and the cognitive state of the observer (estimator variables), these studies do so using a real-world context that highlights relevance to the courts. Applied research on other eyewitness variables, such as those associated with the type of lineup used for identification (system variables), has led to novel insights that can be used to improve eyewitness performance.

Performance Measures

The eyewitness identification problem studied under known-source laboratory conditions is a two-choice classification task, in which the witness subject must report whether a suspect is, or is not, the person they viewed at the crime scene. A few words about assessment of performance on this type of task are useful here.

Discrimination

The ability of an eyewitness to discriminate the perpetrator from innocent suspects is a core question in many laboratory studies, as it can be used to evaluate the impact of estimator and system variables on eyewitness performance. The standard measure of discriminability used for this purpose today is assessed as the frequency with which a witness identifies the perpetrator, relative to the frequency with which the witness identifies an innocent suspect, integrated across all decision criteria, ranging from circumspect to reckless.¹⁰⁷

The knowledge gained from laboratory studies that employ this discriminability measure is an indispensable basis for policy decisions about the best lineup procedures. The level of discriminability associated with different lineup procedures can also be used to infer their relative effectiveness at yielding accurate testimony. This approach reveals nothing, however, about the probability

107. This has long been an effective tool used in basic research on the performance of biological systems for sensation and memory. See D. M. Green & J. A. Swets, *Signal Detection Theory and Psychophysics* (1966).

that a specific witness identification is correct, which is of course what the trier of fact really needs to know.¹⁰⁸

Accuracy

Probability of correctness, or “accuracy,” is distinct from discriminability and is defined in this context as the ratio of correct identifications of the perpetrator relative to all (correct or incorrect) identifications reported. Accuracy of a population of subjects is a simple thing to measure in laboratory studies because ground truth is known, and many recent studies have done so. Like discriminability, accuracy can be used to assess the diagnostic utility of evidence under specified conditions. The important question here is whether it is possible to predict accuracy in real criminal cases, where ground truth is not known. Accuracy prediction in this case must be based on other variables that have been shown to correlate with accuracy, such as DNA genotyping.

Estimator Variables

The states of estimator variables reflect viewing conditions and cognitive status of the witness. Although not controllable by the criminal justice system, these states can be quantified for a given instance and used to predict the fidelity of vision and memory. In what follows, we examine several estimator variables that are commonly associated with real criminal cases in order to illustrate the range of effects on performance as demonstrated in laboratory studies.

Weapon Focus

“Weapon focus” is a phenomenon in which a witness’s focus of visual attention is directed away from the object that may be the most informative in the long run—the face of the perpetrator—and toward the object that is perceived as most immediately threatening to life.¹⁰⁹ As we have learned from basic research on vision (see section titled “Visual attention” above), attention is a limited resource that can be hijacked by compelling features in a visual scene. Just as a driver must focus attention on oncoming traffic at the expense of enjoying scenery along the road, a witness’s life may depend upon the ability to keep attention focused on the perpetrator’s

108. David Dunning & Liza Beth Stern, *Distinguishing Accurate from Inaccurate Eyewitness Identifications via Inquiries About Decision Processes*, 67 J. Personality & Soc. Psych. 818 (1994), <https://doi.org/10.1037//0022-3514.67.5.818>.

109. Elizabeth F. Loftus, *Eyewitness Testimony* (1979).

handgun at the expense of gathering precise information about the perpetrator's face.¹¹⁰

To get some sense of the prevalence and magnitude of the weapon focus effect in the eyewitness context, many applied studies have manipulated the presence of a weapon and measured eyewitness performance.¹¹¹ The consensus from well-designed studies is precisely what we should expect from the basic science of visual attention and inattention blindness: Brandishing a weapon limits the ability of eyewitnesses to discriminate between perpetrators and innocent suspects,¹¹² and markedly impairs identification accuracy.¹¹³

Laboratory studies of weapon focus may significantly underestimate the magnitude of the effect in actual crimes, since the weapon has little threat value in a laboratory setting. Some efforts have been made to address this concern by correlating the presence of a weapon in actual crimes to identification accuracy,¹¹⁴ but interpretation of these analyses is hampered by lack of experimental control over other variables and questionable assumptions of ground truth.¹¹⁵

Finally, because the weapon focus effect is an attention-based phenomenon, it bears on the utility of the *Manson* assertion that eyewitness reliability (the probability that the testimony is correct) can be predicted by “the witness’ degree of attention.” As reviewed above, basic research demonstrates that there are two components to attention: magnitude (or “degree”) and directional focus. Magnitude of attention is indeed high under conditions in which a weapon is present, but the directional focus of attention in such cases precludes accurate identification of the perpetrator, because the eyewitness’s attention is directed at the weapon and not the perpetrator’s face.

Cross-Race Effect

The cross-race effect is a well-known phenomenon of visual cognition and perceptual learning, manifested as better performance on a task that involves

110. Loftus et al., *supra* note 62.

111. Nancy Mehrkens Steblay, *A Meta-Analytic Review of the Weapon Focus Effect*, 16 Law & Hum. Behav. 413 (1992), <https://doi.org/10.1007/bf02352267>; Jonathan M. Fawcett et al., *Of Guns and Geese: A Meta-Analytic Review of the ‘Weapon Focus’ Literature*, 19 Psych., Crim. & Law 35 (2011), <https://doi.org/10.1080/1068316x.2011.599325>.

112. Curt A. Carlson & Maria A. Carlson, *An Evaluation of Lineup Presentation, Weapon Presence, and a Distinctive Feature Using ROC Analysis*, 3 J. Applied Rsch. Memory & Cognition 45 (2014), <https://doi.org/10.1016/j.jarmac.2014.03.004>.

113. Curt A. Carlson & Maria A. Carlson, *A Distinctiveness-Driven Reversal of the Weapon-Focus Effect*, 6(1) Applied Psych. Crim. Just. 82 (2012), <https://doi.org/10.1037/h0101806>.

114. Fawcett et al., *supra* note 111, at 1–32.

115. *Identifying the Culprit*, *supra* note 12.

discriminating between faces of one's own race relative to faces of a different race.¹¹⁶ The effect has been studied extensively in basic research on vision and memory (see sections titled “How Vision Works” and “How Memory Works” above) and is, in part, a by-product of categorical perception. The strength and context generality of this perceptual phenomenon offer good reasons to suspect that it will adversely impact the accuracy of eyewitness identification. Many laboratory studies of eyewitness identification over the past 50 years have confirmed this: The ability to discriminate between the culprit and an innocent suspect suffers and the accuracy of identifications is significantly impaired under conditions in which the eyewitness is of a different race from the lineup participants.¹¹⁷ This impaired eyewitness recognition ability has also been demonstrated experimentally in noncriminal real-world contexts,¹¹⁸ suggesting that it is not a laboratory artifact. Moreover, about half of all persons exonerated based on postconviction DNA testing were convicted, in part, based on mistaken cross-race identifications.¹¹⁹ Applied eyewitness studies, together with extensive basic research on this topic, demonstrate that eyewitness testimony employed by police investigators and courts is, on average, less likely to be accurate if the witness and the suspect are of different races.

Viewing Distance, Lighting, and Exposure Duration

These three factors routinely affect the fidelity of visual experience, as revealed by basic research on the operating characteristics of human vision (see section titled “How Vision Works” above). In simple terms that comport with normal human experience, far viewing distances, dim lighting, and short viewing durations all limit the amount of visual information available for object recognition.

116. McGugin et al., *supra* note 15; Wong et al., *supra* note 15.

117. Christian A. Meissner & John C. Brigham, *Thirty Years of Investigating the Own-Race Bias in Memory for Faces: A Meta-Analytic Review*, 7 Psych., Pub. Pol’y & L. 3 (2001), <https://doi.org/10.1037/1076-8971.7.1.3>; Jungwon Lee & Steven D. Penrod, *Three-Level Meta-Analysis of the Other-Race Bias in Facial Identification*, 36 Applied Cognitive Psych. 1106 (2022), <https://doi.org/10.1002/acp.3997>; Jacqueline Katzman & Margaret Bull Kovera, *Potential Causes of Racial Disparities in Wrongful Convictions Based on Mistaken Identifications: Own-Race Bias and Differences in Evidence-Based Suspicion*, 47 Law & Hum. Behav. 23 (2023), <https://doi.org/10.1037/lhb0000503.supp>; Rachel A. Wilson & Tammy L. Sonnentag, *The Effect of Race of the Perpetrator and Misinformation on Eyewitness Accuracy and Confidence*, 26 Mod. Psych. Studs. 8 (2021); Jesse N. Rothweiler, Kerri A. Goodwin & Jeff Kukucka, *Presence of Administrators Differentially Impacts Eyewitness Discriminability for Same-And Other-Race Identifications*, 34 Applied Cognitive Psych. 1530 (2020), <https://doi.org/10.1002/acp.3733>; Margaret Bull Kovera & Andrew J. Evelo, *Eyewitness Identification in its Social Context*, 10 J. Applied Rsch. Memory & Cognition 313 (2021), <https://doi.org/10.1016/j.jarmac.2021.04.003>.

118. John C. Brigham et al., *Accuracy of Eyewitness Identification in a Field Setting*, 42 J. Pers. & Soc. Psych. 673 (1982), <https://doi.org/10.1037//0022-3514.42.4.673>.

119. Garrett, *supra* note 2 at 73.

Unsurprisingly, they all affect eyewitness performance in predictable ways, as revealed by applied studies of eyewitness identification.

For example, one study found that it is harder to identify faces at a distance because of limits on spatial acuity.¹²⁰ Numerous other applied eyewitness studies similarly support the conclusion that as viewing distance increases, eyewitness performance worsens.¹²¹ One study made the claim that eyewitness accuracy declines by approximately half when viewing distance increases from 3 to 20 meters.¹²² The quantitative characteristics of this relationship (and those drawn simply from basic research on vision) suggest that reliable measures of viewing distance in a given case can help predict the accuracy of an identification.

There is also a large basic science literature on visibility of objects as a function of ambient illumination (see Figure 1), which naturally suggests that eyewitness performance should decline as lighting dims. Indeed it does.¹²³ Other studies have looked at the combined effects of distance and lighting, with one report offering a rule of thumb: Eyewitness performance is acceptable only if lighting exceeds 15 lux (roughly twilight) and distance is less than 15 meters.¹²⁴ Findings from these applied studies, in conjunction with known operating characteristics of human vision, can support quantitative estimates of the accuracy of eyewitness reports.

Finally, basic sciences have long established that duration of exposure (viewing time) to a visual stimulus is correlated with information gain.¹²⁵ Much of the impaired performance associated with short viewing durations is simply due to

120. Geoffrey R. Loftus & Erin M. Harley, *Why Is It Easier to Identify Someone Close than Far Away?*, 12 *Psychonomic Bull. & Rev.* 43 (2005), <https://doi.org/10.3758/bf03196348>.

121. Robert F. Lockamy et al., *The Effect of Viewing Distance on Empirical Discriminability and the Confidence–Accuracy Relationship for Eyewitness Identification*, 34 *Applied Cognitive Psych.* 1047 (2020), <https://doi.org/10.1002/acp.3683>; James Michael Lampinen et al., *Effects of Distance on Face Recognition: Implications for Eyewitness Identification*, 21 *Psychonomic Bull. & Rev.* 1489 (2014), <https://doi.org/10.3758/s13423-014-0641-2>; C.L. Lindsay et al., *How Variations in Distance Affect Eyewitness Reports and Identification Accuracy*, 32 *Law & Hum. Behav.* 526 (2008), <https://doi.org/10.1007/s10979-008-9128-x>; Thomas J. Nyman et al., *The Distance Threshold of Reliable Eyewitness Identification*, 43 *Law & Hum. Behav.* 527 (2019), <https://doi.org/10.1037/lhb0000342.supp>; Thomas J. Nyman et al., *A Stab in the Dark: The Distance Threshold of Target Identification in Low Light*, 6 *Cogent Psych.* 1632047 (2019), <https://doi.org/10.1080/23311908.2019.1632047>.

122. Lockamy et al., *supra* note 63.

123. Lisa DiNardo & David Rainey, *The Effects of Illumination Level and Exposure Time on Facial Recognition*, 41 *Psych. Rec.* 329 (1991), <https://doi.org/10.1007/bf03395115>; Marloes De Jong et al., *Familiar Face Recognition as a Function of Distance and Illumination: A Practical Tool for Use in the Courtroom*, 11 *Psych., Crim. & L.* 87 (2005), <https://doi.org/10.1080/10683160410001715123>.

124. Willem A. Wagenaar & Juliette H. Van Der Schrier, *Face recognition as a function of distance and illumination: A practical tool for use in the courtroom*, 2 *Psych., Crim. & L.* 321 (1996), <https://doi.org/10.1080/10683169608409787>.

125. George Sperling, *The Information Available in Brief Visual Presentations*, 74 *Psych. Monographs: Gen. & Applied* 1 (1960), <https://doi.org/10.1037/h0093759>.

loss of acuity and reduced contrast sensitivity.¹²⁶ Sufficiently long viewing times, by contrast, equip observers with the information needed to make better decisions about what they have seen.¹²⁷ Viewing durations are, of course, not under experimental control in actual eyewitness cases and, regrettably, they are not often recorded with precision,¹²⁸ but one report estimates that over a third of all eyewitness reports are based on viewing durations of less than one minute.¹²⁹ Numerous laboratory studies of eyewitness performance have manipulated viewing time and found, as expected, that longer viewing times are associated with greater eyewitness accuracy,¹³⁰ although high-confidence identifications are generally highly accurate for both short and long exposure durations,¹³¹ an issue we take up in the section titled “Confidence” below.

126. Jacob Nachmias, *Effect of Exposure Duration on Visual Contrast Sensitivity with Square-Wave Gratings*, 57 J. Opt. Soc. Am. 421 (1967), <https://doi.org/10.1364/josa.57.000421>; J. Jason McAnany, *The Effect of Exposure Duration on Visual Acuity for Letter Optotypes and Gratings*, 105 Vision Rsch. 86 (2014), <https://doi.org/10.1016/j.visres.2014.08.024>.

127. Radoslaw Martin Cichy, Dimitrios Pantazis & Aude Oliva, *Resolving Human Object Recognition in Space and Time*, 17 Nature Neurosci. 455 (2014), <https://doi.org/10.1038/nn.3635>.

128. An eyewitness may be the only source of information regarding the duration of exposure to the perpetrator during a crime. Evidence suggests that these estimates are rarely accurate. Holly L. Gasper, Michael M. Roy & Heather D. Flowe, *Improving Time Estimation in Witness Memory*, 10 Frontiers Psych. 1452 (2019), <https://doi.org/10.3389/fpsyg.2019.01452>; A. Daniel Yarmey & Meagan J. Yarmey, *Eyewitness Recall and Duration Estimates in Field Settings*, 27 J. Applied Soc. Psych. 330 (1997), <https://doi.org/10.1111/j.1559-1816.1997.tb00635.x>.

129. Tim Valentine, Alan Pickering & Stephen Darling, *Characteristics of Eyewitness Identification That Predict the Outcome of Real Lineups*, 17 Applied Cognitive Psych. 969 (2003), <https://doi.org/10.1002/acp.939>.

130. J. Kirkland Reynolds & Kathy Pezdek, *Face Recognition Memory: The Effects of Exposure Duration and Encoding Instruction*, 6 Applied Cognitive Psych. 279 (1992), <https://doi.org/10.1002/acp.2350060402>; Amina Memon, Lorraine Hope & Ray Bull, *Exposure Duration: Effects on Eyewitness Accuracy and Confidence*, 94 Brit. J. Psych. 339 (2003), <https://doi.org/10.1348/000712603767876262>; Matthew A. Palmer et al., *The Confidence-Accuracy Relationship for Eyewitness Identification Decisions: Effects of Exposure Duration, Retention Interval, and Divided Attention*, 19 J. Exper. Psych.: Applied 55 (2013), <https://doi.org/10.1037/a0031602>; Rolando N. Carol & Nadja Schreiber Compo, *The Effect of Encoding Duration on Implicit and Explicit Eyewitness Memory*, 61 Consciousness & Cognition 117 (2018), <https://doi.org/10.1016/j.concog.2018.02.004>; Brian H. Bornstein et al., *Effects of Exposure Time and Cognitive Operations on Facial Identification Accuracy: A Meta-Analysis of Two Variables Associated with Initial Memory Strength*, 18 Psych., Crim. & L. 473 (2012), <https://doi.org/10.1080/1068316x.2010.508458>; Neil Brewer, Michael Gordon & Nigel Bond, *Effect of Photoarray Exposure Duration on Eyewitness Identification Accuracy and Processing Strategy*, 6 Psych., Crim. & L. 21 (2000), <https://doi.org/10.1080/10683160008410829>.

131. Carolyn Semmler et al., *The Role of Estimator Variables in Eyewitness Identification*, 24 J. Exper. Psych.: Applied 400 (2018), <https://doi.org/10.1037/xap0000157>; Matthew A. Palmer et al., *The Confidence-Accuracy Relationship for Eyewitness Identification Decisions: Effects of Exposure Duration, Retention Interval, and Divided Attention*, 19 J. Exper. Psych.: Applied 55 (2013), <https://doi.org/10.1037/a0031602.supp>.

The voluminous body of data on exposure duration thus demonstrates that, when reliable information about viewing times is available, this information can be used to infer the probability that eyewitness testimony is accurate.

Multiple Perpetrators

Crimes are often carried out by more than one person. Evidence indicates that the number of multiperpetrator homicides, for example, has been on the rise for decades and is estimated to exceed 20% of all murders.¹³² Explanations point to growing gang violence, peer pressure, and the perception of distributed responsibility (“de-individuation” of any single perpetrator)—a form of groupthink that licenses behaviors that individuals acting alone would view as impermissible.

For our purposes, the important empirical question is: How well can a witness be expected to remember one or more faces under these circumstances? This is a divided attention problem, much like the weapon focus problem addressed above, in which limited attentional resources must be distributed across multiple targets. While focused attention generally improves the fidelity of vision and memory for an attended stimulus, considerable evidence from basic science demonstrates what comports with everyday human experience: Detection, discriminability, and recognition of visual targets is impaired when attention must be subdivided between them.¹³³ A number of laboratory studies have explored the consequences of this phenomenon for eyewitness identification.¹³⁴ As expected (and mirroring the weapon focus effect on divided attention), identification accuracy is consistently poorer when multiple perpetrators are

132. Lockamy et al., *supra* note 63; Nortje et al., *supra* note 63, at 348–82.

133. Edward Awh & Harold Pashler, *Evidence for Split Attentional Foci*, 26 J. Exper. Psych. Hum. Perception & Performance 834 (2000), <https://doi.org/10.1037//0096-1523.26.2.834>; Arthur F. Kramer & Sowon Hahn, *Splitting the Beam: Distribution of Attention over Noncontiguous Regions of the Visual Field*, 6 Psych. Sci. 381 (1995), <https://doi.org/10.1111/j.1467-9280.1995.tb00530.x>; Won Mok Shim, George A. Alvarez & Yuhong V. Jiang, *Spatial Separation Between Targets Constrains Maintenance of Attention on Multiple Objects*, 15 Psychonomic Bull. & Rev. 390 (2008), <https://doi.org/10.3758/pbr.15.2.390>; Steven Yantis, *Multi-Element Visual Tracking: Attention and Perceptual Organization*, 24 Cognitive Psych. 295 (1992), [https://doi.org/10.1016/0010-0285\(92\)90010-y](https://doi.org/10.1016/0010-0285(92)90010-y); Amelia H. Harrison, Sam Ling & Joshua J. Foster, *The Cost of Divided Attention for Detection of Simple Visual Features Primarily Reflects Limits in Post-Perceptual Processing*, 85 Attention, Perception, & Psychophysics 377 (2023), <https://doi.org/10.3758/s13414-022-02547-7>; Harold Pashler, *Attention and Visual Perception: Analyzing Divided Attention*, in *Visual Cognition: An Invitation to Cognitive Science* 71 (2d ed. 1995), <https://doi.org/10.7551/mitpress/3965.003.0005>; M. Corbetta et al., *Selective and Divided Attention During Visual Discriminations of Shape, Color, and Speed: Functional Anatomy by Positron Emission Tomography*, 11 J. Neurosci. 2383 (1991), <https://doi.org/10.1523/jneurosci.11-08-02383.1991>; Nilli Lavie, *Distracted and Confused? Selective Attention Under Load*, 9 Trends Cognitive Sci. 75 (2005), <https://doi.org/10.1016/j.tics.2004.12.004>.

134. Clifford & Hollin, *supra* note 63; Hobson & Wilcock, *supra* note 63; Lockamy et al., *supra* note 63.

witnessed during a crime. For the purposes of the courts, it follows that eyewitness testimony under these conditions is less likely to be correct.

Facial Recognition Ability

Basic research shows (see section titled “Categorical perception” above) that one feature of categorical visual perception is category-specific expertise. While humans have a highly specialized system for face processing,¹³⁵ face recognition ability varies significantly across the population, from “super recognizers” at one extreme to “face-blind people” at the other (see Figure 12).¹³⁶ Chad Dodson and colleagues hypothesized that knowledge of the facial recognition ability of a witness might provide insight into the probative value of an identification by that witness.¹³⁷ Results from this laboratory study reveal that a high-confidence identification is far more likely to be in error if the witness scores poorly on a test of facial recognition ability. These findings indicate that a simple test of face recognition ability can provide factfinders with valuable insight into the accuracy of high confidence identifications. We would not hesitate to conduct a visual acuity test to explore whether poor eyesight could have affected an eyewitness identification. Similarly, testing the face recognition ability of an eyewitness could become a standard procedure for law enforcement and for courts, which serves the purpose of inferring identification accuracy.

Retention Interval

As noted above (see section titled “Memory Storage”), many decades of basic research on memory have revealed that stored information is lost with the passage of time, as evidenced by progressive performance deterioration on standard visual discrimination and recognition tasks (see Figure 11). This memory loss is most rapid immediately after acquisition, when consolidation into long-term memory is not yet complete. After this initial period, consolidation continues but resulting long-term memories are readily modified by exposure to new information, the likelihood of which naturally increases with time. A witness’s inevitable interactions with law enforcement and legal counsel, not to mention communications from journalists, family, and friends, have the potential to significantly modify the witness’s memory of faces encountered and other details of

135. Winrich Freiwald, Bradley Duchaine & Galit Yovel, *Face Processing Systems: From Neurons to Real World Social Perception*, 39 Ann. Rev. Neurosci. 325 (2016), <https://doi.org/10.1146/annurev-neuro-070815-013934>.

136. Duchaine & Nakayama, *supra* note 101.

137. Grabman et al., *supra* note 103, at 233–43.

the crime scene.¹³⁸ Thus, the fidelity of retrieved events—and the accuracy of identification—is likely to be greater when retrieval occurs closer to the time of the witnessed events.

These concerns about the completeness and stability of long-term memories have motivated a large number of applied studies of performance on eyewitness identification as a function of retention interval, defined as the length of time from the witnessed event to the lineup. A meta-analysis revealed that performance declines significantly as retention interval increases.¹³⁹ Moreover, the authors maintain that the rate of memory loss, modeled as a power law,¹⁴⁰ can be used to estimate the probability that an identification is correct: “[N]ot only can the percentage of remaining memory strength be determined for any retention interval, but this strength estimate can be translated into an estimated probability of being correct on a fair lineup of a specified size.”¹⁴¹

The initial strength of an encoded face memory is, of course, modulated by other estimator variables associated with witnessing the crime, such as cognitive state (e.g., attention and emotion), viewing conditions (e.g., distance, lighting, and exposure duration), and the cross-race effect.¹⁴² Nonetheless, there are strategies for estimating initial memory strength, and one study makes the promising claim that “we can now offer the judge or juror an estimate of what proportion of memory strength, regardless of its initial value, remained at the time the eyewitness’s memory for the perpetrator was tested.”¹⁴³ This predictive approach could provide a quantitative foundation for the fifth *Manson* criterion for accuracy prediction: “the length of time between the crime and the confrontation.”

Stress and Emotion

As noted above in our review of the basic sciences of vision and memory (see sections titled “Visual attention” and “The effect of emotion on memory storage” above), considerable empirical evidence demonstrates that stress and high states of

138. Maria S. Zaragoza & Sean M. Lane, *Source Misattributions and the Suggestibility of Eyewitness Memory*, 20 J. Exper. Psych.: Learning, Memory & Cognition 934 (1994), <https://doi.org/10.1037//0278-7393.20.4.934>; William C. Thompson, et al., *What Did the Janitor Do? Suggestive Interviewing and the Accuracy of Children’s Accounts*, 21 Law & Hum. Behav. 405 (1997), <https://doi.org/10.1023/a:1024859219764>; D. Stephen Lindsay & Marcia K. Johnson, *The Eyewitness Suggestibility Effect and Memory for Source*, 17 Memory & Cognition 349 (1989), <https://doi.org/10.3758/bf03198473>.

139. Deffenbacher et al. *supra* note 26, at 139.

140. Wickelgren, *supra* note 85; Wixted & Ebbesen, *supra* note 85.

141. Deffenbacher et al., *supra* note 26.

142. Melissa A. Pigott, John C. Brigham & Robert K. Bothwell, *A Field Study on the Relationship Between Quality of Eyewitnesses’ Descriptions and Identification Accuracy*, 17 J. Police Sci. & Admin. 84 (1990), <https://perma.cc/9GXP-8DWH>.

143. Deffenbacher et al., *supra* note 26.

emotion can influence the allocation of visual attention as well as the encoding and retrieval of memories. Witnessing a crime, and the subsequent identification event, frequently elicit stress and often trigger high levels of emotion (commonly fear)—particularly in cases where the witness is a victim, a weapon is involved, or there is serious threat to life or limb. It naturally follows that these altered states of the observer may affect the content and validity of eyewitness reports.

The effects of stress and emotion on eyewitness performance have been difficult to study, as levels that often accompany eyewitness events in the real world are extremely difficult to replicate in a laboratory setting, in part because of the obvious pretense and the fact that regulations prohibit induction of high stress levels in human subjects. Some studies have, nonetheless, reportedly achieved moderate levels of stress and fear using unusual techniques, which generally involve manipulations that are ancillary to the identification test itself. One experiment involved subjecting active-duty military personnel to a mock POW camp (a valid part of military survival school training), which included aggressive interrogation and physical isolation-related stress.¹⁴⁴ The study found that memories acquired during stressful events are highly vulnerable to modification by exposure to postevent misinformation, even in individuals whose level of training and experience might be considered relatively immune to such influences. Other studies of this sort have attempted to elicit stress and fear by placing subjects serving as witnesses in a creepy environment (the Horror Labyrinth of the London Dungeon)¹⁴⁵ or administering vaccine injections to children who were serving as witnesses.¹⁴⁶

Another approach in laboratory studies has been to independently confirm that “putative manipulations of stress/anxiety or perceived level of violence had been successful.”¹⁴⁷ One meta-analysis limited to studies that obtained such confirmation (using standard physiological measures, such as breathing, and heart rate) found, as expected, that confirmed states of stress were correlated with impairments of identification accuracy and impaired ability to identify key characteristics of a witnessed individual’s face (e.g., hair length, hair color, eye color, shape of face, presence of facial hair). As emphasized in a more recent review, “ensuring valid stress inductions is a prerequisite for effectively studying effects of acute stress on memory performance.”¹⁴⁸

144. C.A. Morgan III et al., *Misinformation Can Influence Memory for Recently Experienced Highly Stressful Events*, 36 Int’l J.L. & Psychiatry 11 (2013), <https://doi.org/10.1016/j.ijlp.2012.11.002>.

145. Tim Valentine & Jan Mesout, *Eyewitness Identification Under Stress in the London Dungeon*, 23 Applied Cognitive Psych. 151 (2009), <https://doi.org/10.1002/acp.1510>.

146. Douglas P. Peters, *Stress, Arousal, and Children’s Eyewitness Memory*, in *Memory for Everyday and Emotional Events* 351 (1st ed. 2018), <https://doi.org/10.4324/9781315789231-15>.

147. Kenneth A. Deffenbacher et al., *A Meta-Analytic Review of the Effects of High Stress on Eyewitness Memory*, 28 Law & Hum. Behav. 687 (2004), <https://doi.org/10.1007/s10979-004-0565-x>.

148. Carey Marr et al., *The Effects of Acute Stress on Eyewitness Memory: An Integrative Review for Eyewitness Researchers*, 29 Memory 1091 (2021), <https://doi.org/10.1080/09658211.2021.1955935>.

Laboratory studies have also revealed a counterintuitive relationship between accuracy and confidence under conditions of stress. Specifically, while stress clearly interferes with the accuracy of eyewitness memory, the predictive relationship between confidence and accuracy is similar across stressful and stress-free viewing conditions. (Stress-free observers are highly confident when accurate, while stressed observers are not confident when inaccurate.) This stability of the confidence–accuracy relationship implies that stressed witnesses are sufficiently aware of their cognitive state, and its potential to limit accuracy, and they “scale back their confidence judgments” accordingly.¹⁴⁹ These findings imply that confidence reports may be useful predictors of accuracy under conditions of stress.

Despite the promise of these various approaches, it is impossible under laboratory conditions to achieve levels of stress and fear experienced by, for example, real witness–victims of rape, or a witness to the murder of a relative or friend, or a mass shooting. Nonetheless, although *Manson* made no mention of stress or fear in their predictive framework, the existing data argue that if stress or fear can be inferred from the circumstances of a witnessed crime, any eyewitness testimony proffered to the court will be of questionable accuracy.

Confidence

No topic in the literature on eyewitness identification has elicited as much robust discussion and research as confidence and its relationship to accuracy.¹⁵⁰ In broad terms, this debate stems from the conflicting goals of (a) finding a measure that can predict the accuracy of an identification, and (b) avoiding the enormous personal and societal risk that such a measure might fail.

Before we review the basis for this debate, it is useful to define what we mean by confidence in this context and illustrate how it works. Confidence has a broad colloquial definition, but in the eyewitness context it refers to the perceived degree of certainty in a piece of information or a decision. Confidence normally grows when multiple sources of information paint a coherent picture.¹⁵¹ Many decades of basic research on confidence and the accuracy of

149. Sara D. Davis et al., *Physiological Stress and Face Recognition: Differential Effects of Stress on Accuracy and the Confidence–Accuracy Relationship*, 8 J. Applied Rsch. Memory & Cognition 367 (2019), <https://doi.org/10.1016/j.jarmac.2019.05.006>.

150. Confidence has traditionally been treated as an estimator variable in applied eyewitness research, since it is a property of the witness under the conditions of viewing. As this science has advanced, however, it has become clear that confidence as a variable is not so easily pigeonholed: It can be predictably influenced by the state of systems variables.

151. Kahneman & Tversky, *supra* note 77, at 237–51; Daniel Kahneman, *Thinking, Fast and Slow* (2011).

retrieved memories reveal that the correlation is frequently poor,¹⁵² which we can largely attribute to the natural process of smoothing over source discrepancies to create a sense of coherence, which, in turn, fosters confidence. This dissociation between confidence and accuracy has obvious implications for eyewitness testimony, and many have warned of the risks associated with the use of witness confidence statements for criminal investigation and prosecution.¹⁵³

Timing of identification and conditions of confidence assessment

Applied eyewitness research has revealed nuances of the confidence–accuracy relationship. For example, the problem of confidence inflation,¹⁵⁴ in which witnesses progressively gain greater confidence during the window of time between the initial lineup identification and the courtroom identification, has led to the recommendation that only the initial identification and confidence statements made at that time be used as evidence.¹⁵⁵ Considerable evidence supports this

152. Loftus, *supra* note 109; Loftus & Pickrell, *supra* note 88, at 720–25; Kenneth A. Norman & Daniel L. Schacter, *False Recognition in Younger and Older Adults: Exploring the Characteristics of Illusory Memories*, 25 *Memory & Cognition* 838 (1997), <https://doi.org/10.3758/bf03211328>; Henry L. Roediger III & Kathleen B. McDermott, *Creating False Memories: Remembering Words Not Presented in Lists*, 21 *J. Exper. Psych.: Learning, Memory & Cognition* 803 (1995), <https://doi.org/10.1037//0278-7393.21.4.803>; Thomas A. Busey et al., *Accounts of the Confidence–Accuracy Relation in Recognition Memory*, 7 *Psychonomic Bull. & Rev.* 26 (2000) <https://doi.org/10.3758/BF03210724>; Daniel L. Schacter & Chad S. Dodson, *Misattribution, False Recognition and the Sins of Memory*, 356 *Phil. Transactions Royal Soc. London Series B: Biological Scis.* 1385 (2001), <https://doi.org/10.1098/rstb.2001.0938>; Hal R. Arkes, *Overconfidence in Judgmental Forecasting*, in *Principles of Forecasting* 495 (2001), https://doi.org/10.1007/978-0-306-47630-3_22; Mark Tippens Reinitz et al., *Different Confidence–Accuracy Relationships for Feature-Based and Familiarity-Based Memories*, 37 *J. Exper. Psych.: Learning, Memory & Cognition* 507 (2011), <https://doi.org/10.1037/a0021961>.supp; Henry L. Roediger III, John H. Wixted & K. Andrew DeSoto, *The Curious Complexity between Confidence and Accuracy in Reports from Memory*, in *Memory and Law* 84–118 (Lynn Nadel & Walter Sinnott-Armstrong eds., 2012); Oxford University Press, New York; Elizabeth F. Loftus & Rachel L. Greenspan, *If I’m Certain, Is It True? Accuracy and Confidence in Eyewitness Memory*, 18 *Psych. Sci. Pub. Int.* 1 (2017), <https://doi.org/10.1177/1529100617699241>.

153. Sporer et al., *supra* note 75; Deffenbacher, *supra* note 75; Reinitz et al., *supra* note 75; Bothwell et al., *supra* note 75.

154. Melissa Paiva et al., *Influence of Confidence Inflation and Explanations for Changes in Confidence on Evaluations of Eyewitness Identification Accuracy*, 16 *Legal & Criminological Psych.* 266 (2011), <https://doi.org/10.1348/135532510x503340>; Sporer et al., *supra* note 75; Deffenbacher, *supra* note 75; Reinitz et al., *supra* note 75; Bothwell et al., *supra* note 75; Laura Smalarz & Gary L. Wells, *Do Multiple Doses of Feedback Have Cumulative Effects on Eyewitness Confidence?*, 9 *J. Applied Rsch., Memory & Cognition* 508 (2020), <https://doi.org/10.1037/h0101857>; Nancy K. Steblay, Gary L. Wells & Amy Bradfield Douglass, *The Eyewitness Post Identification Feedback Effect 15 Years Later: Theoretical and Policy Implications*, 20 *Psych., Pub. Pol’y & L.* 1 (2014), <https://doi.org/10.1037/law0000001>; Douglass & Steblay, *supra* note 75.

155. Identifying the Culprit, *supra* note 12.

recommendation: Memory is patently malleable,¹⁵⁶ and the very act of probing memory in a lineup procedure can actually modify that memory in a way that will lessen the accuracy of future identifications.¹⁵⁷ Following this logic, an argument can be made that the initial identification should be the *only* identification obtained.¹⁵⁸

Additional insight into this issue has come from work by John Wixted and colleagues, who argued that confidence reports made at the time of the initial identification have their greatest predictive value if they are held to a very high standard. Specifically, these investigators report that the relationship between confidence and accuracy is strong if the initial identification is made under “pristine” conditions.¹⁵⁹ Those conditions include a properly blinded test and fair lineup construction (no participant stands out), which promote coherence of information received by the eyewitness, thus limiting the accuracy-reducing inclination to smooth over discrepancies.

Critics argue that it may be difficult to recognize pristine conditions in real casework, owing in part to limited access to situational context (e.g., what exactly did the police do?). By this argument, eyewitness confidence may place innocent suspects at great risk.¹⁶⁰ Some disagreement remains on this point,¹⁶¹ but the rationale for the pristine distinction and the use of only the initial identification and confidence statement rests firmly on the cognitive science of memory and decision-making.¹⁶²

156. Deborah Davis & Elizabeth F. Loftus, *Eyewitness Science in the 21st Century: What Do We Know and Where Do We Go from Here?*, in 1 Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience (4th ed. 2018), <https://doi.org/10.1002/9781119170174.epcn116>.

157. Nancy K. Steblay, Robert W. Tix & Samantha L. Benson, *Double Exposure: The Effects of Repeated Identification Lineups on Eyewitness Accuracy*, 27 Applied Cognitive Psych. 644 (2013), <https://doi.org/10.1002/acp.2944>; Nancy K. Steblay & Jennifer E. Dysart, *Repeated Eyewitness Identification Procedures with the Same Suspect*, 5 J. Applied Rsch. Memory & Cognition 284 (2016), <https://doi.org/10.1016/j.jarmac.2016.06.010>.

158. John T. Wixted et al., *Test a Witness's Memory of a Suspect Only Once*, 22 Psych. Sci. Pub. Int. 1S (2021), <https://doi.org/10.1177/15291006211026259>.

159. *Id.*; Wixted & Wells, *supra* note 25; Semmler et al., *supra* note 131.

160. Shari R. Berkowitz & Steven J. Frenda, *Rethinking the Confident Eyewitness: A Reply to Wixted, Mickes, and Fisher*, 13 Persps. Psych. Sci. 336 (2018), <https://doi.org/10.1177/1745691617751883>; Shari R. Berkowitz et al., *Convicting with Confidence? Why We Should Not Over-Rely on Eyewitness Confidence*, 30 Memory 10 (2022), <https://doi.org/10.1080/09658211.2020.1849308>; Shari R. Berkowitz et al., *Eyewitness Confidence May Not be Ready for the Courts: A Reply to Wixted Et Al.*, 30 Memory 75 (2022), <https://doi.org/10.1080/09658211.2021.1952271>; James D. Sauer, Matthew A. Palmer & Neil Brewer, *Pitfalls in Using Eyewitness Confidence to Diagnose the Accuracy of an Individual Identification Decision*, 25 Psych., Pub. Pol'y & L. 147 (2019), <https://doi.org/10.1037/law0000203>.

161. Travis M. Seale-Carlisle et al., *Confidence and Response Time as Indicators of Eyewitness Identification Accuracy in the Lab and in the Real World*, 84 J. Applied Rsch. Memory & Cognition 420 (2019), <https://doi.org/10.1016/j.jarmac.2019.09.003>; Wixted & Wells, *supra* note 25.

162. Semmler et al., *supra* note 131.

System Variables

System variables are causal states that are potentially under the control of the criminal justice system, since they bear on activities that occur after the witnessed events. These include (1) communication between law enforcement and the witness, (2) communication between the witness and the larger community (legal defense, media, family and friends, etc.), (3) the practice of “evidence-based suspicion,” (4) prior exposure to images of potential suspects, (5) the type of procedure used for identification, and (6) the choice of lineup fillers (lineup participants known to be innocent).

The empirical findings and recommendations derived from research on system variables are of direct relevance to, and advisory of, police practices. That is, they can be used to improve eyewitness identification performance. Knowledge of the states of these variables for a given case, however, can also be useful to the courts in estimating the probability that eyewitness testimony gained by law enforcement is accurate.

Communication Between Law Enforcement Personnel and Witness

Statements that convey prior probability of suspicion

Police commonly have information that could influence both the accuracy of a witness’s identification and the confidence expressed in the identification. For example, a police officer might first report to the witness/victim of a carjacking that a suspect was apprehended in the stolen vehicle and then ask the witness to view a lineup with that suspect in it. This might seem innocuous on the surface and could be viewed as reinforcing information for the witness, but it establishes a significant prior probability that the perpetrator is in the lineup. Inferences drawn from such priors can readily bias perception, choice, and action under conditions of uncertainty.¹⁶³ The result is that the witness may actually *perceive* the perpetrator in a target-absent lineup and may unwittingly adopt a lower criterion for reporting an identification even when recognition memory is poor.

Communication during lineups: The importance of blinded lineup administration

In cases where the administrator of a lineup knows the status of lineup participants, they are in a position to inadvertently communicate that information to a

163. Albright, *supra* note 36.

witness. The witness, commonly eager to be of assistance, can easily pick up unintended signals (e.g., direction of gaze, body posture, facial expressions), which may be highly suggestive and have the effect of biasing the outcome—yielding a larger fraction of misidentifications and inflation of witness confidence. Generally, the solution to these biasing effects is to limit communication between the witness and people with relevant knowledge.¹⁶⁴

Many applied eyewitness studies support this information-control approach and demonstrate that the problem of inadvertent communication during a lineup can be eliminated by the practice of “double blinding.”¹⁶⁵ This simple and effective tool is a pillar of the scientific method.¹⁶⁶ In an experimental test of a novel drug, for example, a subject is prevented (hence “blinded” or “single blinded”) from knowing the experimental condition that they have been assigned to. Furthermore, the scientist who measures the effect of the drug is also prevented (“double blinded”) from knowing the experimental condition associated with that measure, all of which precludes bias based on prior knowledge. Similarly, in a lineup, the witness is naturally blinded to knowledge of the “conditions,” but the risk of bias can be reduced if the lineup administrator has no knowledge to convey—that is, the test is double blinded. This simple practice is highly effective at improving eyewitness accuracy, and the risks of not conducting double-blinded lineups are clear. For all of these reasons, failure to do so runs afoul of

164. Ryann M. Haw & Ronald P. Fisher, *Effects of Administrator-Witness Contact on Eyewitness Identification Accuracy*, 89 J. Applied Psych. 1106 (2004), <https://doi.org/10.1037/0021-9010.89.6.1106>; Robert Rosenthal, *Covert Communication in Classrooms, Clinics, Courtrooms, and Cubicles*, 57 Am. Psych. 839 (2002), <https://doi.org/10.1037//0003-066x.57.11.839>.

165. Steven E. Clark, Tanya E. Marshall & Robert Rosenthal, *Lineup Administrator Influences on Eyewitness Identification Decisions*, 15 J. Exper. Psych.: Applied 63 (2009), <https://doi.org/10.1037/a0015185>; Margaret Bull Kovera & Andrew J. Evelo, *The Case for Double-Blind Lineup Administration*, 23 Psych., Pub. Pol’y & L. 421 (2017), <https://doi.org/10.1037/law0000139>; Margaret Bull Kovera & Andrew J. Evelo, *Improving Eyewitness-Identification Evidence Through Double-Blind Lineup Administration*, 29 Current Directions Psych. Sci. 563 (2020), <https://doi.org/10.1177/0963721420969366>; Steve D. Charman & Vanessa Quiroz, *Blind Sequential Lineup Administration Reduces Both False Identifications and Confidence in Those Identifications*, 40 Law & Hum. Behav. 477 (2016), <https://doi.org/10.1037/lhb0000197>; Dario N. Rodriguez & Melissa A. Berry, *Eyewitness Science and the Call for Double-Blind Lineup Administration*, 2013 J. Criminology 530523 (2013), <https://doi.org/10.1155/2013/530523>; Jacqueline L. Austin et al., *Double-Blind Lineup Administration: Effects of Administrator Knowledge on Eyewitness Decisions*, in 139 Reform of Eyewitness Identification Procedures (2013), <https://doi.org/10.1037/14094-007>; Arthur H. Perlini & Andrew D. Silvaggio, *Eyewitness Misidentification: Single vs. Double-Blind Comparison of Photospread Administration*, 100 Psych. Reps. 247 (2007), <https://doi.org/10.2466/pr0.100.1.247-256>.

166. Michael Stolberg, *Inventing the Randomized Double-Blind Trial: The Nuremberg Salt Test of 1835*, 99 J. Royal Soc. Med. 642 (2006), <https://doi.org/10.1177/014107680609901216>; Lorraine Daston, *Scientific Error and the Ethos of Belief*, 72 Soc. Rsch.: Int’l Q. 1 (2005), <https://doi.org/10.1353/sor.2005.0016>; Paul J. Karanicolas, Forough Farrokhyar & Mohit Bhandari, *Practical Tips for Surgical Research: Blinding: Who, What, When, Why, How?*, 53 Can. J. Surg. 345 (2010); Gary L. Wells & C.A. Elizabeth Luus, *Police Lineups as Experiments: Social Methodology as a Framework for Properly Conducted Lineups*, 16 Pers. & Soc. Psych. Bull. 106 (1990), <https://doi.org/10.1177/0146167290161008>.

best practices,¹⁶⁷ recommendations of the U.S. Department of Justice,¹⁶⁸ and the International Association of Chiefs of Police.¹⁶⁹ In a court of law, lineup blinding status is naturally a useful predictor of eyewitness accuracy.

Video recording of lineup procedures

In view of the risk that some forms of communication between the lineup administrator(s) and witness can bias the identification, many have called for video recording of the entire procedure.¹⁷⁰ One may not expect that doing so will change the behavior of the people involved, but the recording serves as stable evidence to confirm—to the court, to the prosecution, and to the defense—whether the procedure was done in such a way to avoid a biased outcome. It also documents who was present, what the witness actually said about an identification and their level of confidence, and other behavioral characteristics (e.g., time to decide) that may bear on the accuracy of identifications. The absence of such video recordings, conversely, limits inferences about the accuracy of eyewitness testimony.

Postidentification feedback

Postidentification feedback from law enforcement can also influence a witness's confidence in their decision.¹⁷¹ Applied studies have shown that simple reinforcing statements, even without specifics—“Nice work!”—can increase or decrease witness confidence, thus distorting any relationship to accuracy (see section titled “Timing of identification and conditions of confidence assessment” above).¹⁷²

167. Identifying the Culprit, *supra* note 12.

168. U.S. Dep't of Just., Eyewitness Identification, *supra* note 7.

169. IACP (International Association of Chiefs of Police) Law Enforcement Policy Center, Eyewitness Identification: Model Policy Concepts & Issues (2016), <https://perma.cc/3YDT-8A77>.

170. Identifying the Culprit, *supra* note 12; Wells et al., *Policy and Procedure Recommendations* (2020), *supra* note 27, at 3–36. U.S. Dep't of Just., Eyewitness Identification, *supra* note 7.

171. Douglass & Steblay, *supra* note 75; Mitchell L. Eisen et al., *Does Anyone Else Look Familiar? Influencing Identification Decisions by Asking Witnesses to Re-Examine the Lineup*, 42 Law & Hum. Behav. 306 (2018), <https://doi.org/10.1037/lhb0000291.supp>.

172. Carolyn Semmler, Neil Brewer & Gary L. Wells, *Effects of Postidentification Feedback on Eyewitness Identification and Nonidentification Confidence*, 89 J. Applied Psych. 334 (2004), <https://doi.org/10.1037/0021-9010.89.2.334>; Gary L. Wells & Amy L. Bradfield, “Good, You Identified the Suspect”: Feedback to Eyewitnesses Distorts Their Reports of the Witnessing Experience, 83 J. Applied Psych. 360 (1998), <https://doi.org/10.1037//0021-9010.83.3.360>; Amy L. Bradfield, Gary L. Wells & Elizabeth A. Olson, *The Damaging Effect of Confirming Feedback on the Relation Between Eyewitness Certainty and Identification Accuracy*, 87 J. Applied Psych. 112 (2002), <https://doi.org/10.1037//0021-9010.87.1.112>; Nancy K. Steblay, Gary L. Wells & Amy Bradfield Douglass, *The Eyewitness Post*

Very high or very low confidence expressed by a witness can, in turn, be very compelling to the trier of fact, regardless of whether the identification is correct.¹⁷³

Communication Between Community and Witness: Social Influences on Eyewitness Performance

As we have seen, memory is malleable and the criteria that people use for momentous decisions can change over time. People are highly social creatures—eager for social reinforcement and easily swayed by pressure for social conformity—and it should come as no surprise that in the eyewitness context (and more generally), the expectations, real or inferred, of other parties (police, attorneys, news media, family, colleagues and comrades, social media followers) can have a profound influence over what we remember, the decisions we make, and the confidence we express. There are two important features to this potential for bias: the social utility of misinformation and the scale of social networks.

The social utility of misinformation

Social influences can readily modify an individual's memory of events. New information received from one's social networks may be unvetted "misinformation," which is seen as widely accepted by others and thus not held to rigorous standards. People have a regrettable affinity for such information,¹⁷⁴ in part because knowing things, regardless of whether they are true, has social value. The end result is contamination of the original content of memory, such that decisions are less accurate and confidence in them becomes distorted. In the

Identification Feedback Effect 15 Years Later: Theoretical and Policy Implications, 20 Psych., Pub. Pol'y & L. 1 (2014), <https://doi.org/10.1037/law0000001>.

173. Steven D. Penrod & Brian L. Cutler, *Eyewitness Expert Testimony and Jury Decisionmaking*, 52 Law & Contemp. Probs. 43 (1989), <https://doi.org/10.2307/1191907>; Neil Brewer & Anne Burke, *Effects of Testimonial Inconsistencies and Eyewitness Confidence on Mock-Juror Judgments*, 26 Law & Hum. Behav. 353 (2002), <https://doi.org/10.1023/a:1015380522722>; James D. Sauer, Matthew A. Palmer & Neil Brewer, *Mock-Juror Evaluations of Traditional and Ratings-Based Eyewitness Identification Evidence*, 41 Law & Hum. Behav. 375 (2017), <https://doi.org/10.1037/lhb0000235.supp>; Crystal R. Slane & Chad S. Dodson, *Eyewitness Confidence and Mock Juror Decisions of Guilt: A Meta-Analytic Review*, 46 Law & Hum. Behav. 45 (2022), <https://doi.org/10.1037/lhb0000481.supp>.

174. Ilan Yaniv & Dean P. Foster, *Graininess of Judgment Under Uncertainty: An Accuracy-Informativeness Trade-Off*, 124 J. Exper. Psych.: Gen. 424 (1995), <https://doi.org/10.1037/0096-3445.124.4.424>; Rakefet Ackerman & Morris Goldsmith, *Control over Grain Size in Memory Reporting—With and Without Satisficing Knowledge*, 34 J. Exper. Psych.: Learning, Memory & Cognition 1224 (2008), <https://doi.org/10.1037/a0012938>.

extreme, people can hold and act upon “false memories” of things that never happened.¹⁷⁵

Applied studies have explored the influence of postwitnessing information gain on eyewitness accuracy and confidence. The most straightforward of these effects is incorporation of new information into the preexisting memory of the crime scene, without awareness (a form of source memory failure).¹⁷⁶ Acquired knowledge of the eyewitness reports of others, for example, can bring about both a change in a witness’s memory and confidence.¹⁷⁷ Perhaps more pernicious are influences from news reports, social media, and colleagues: Eyewitnesses in laboratory studies have been shown to readily modify their own memories, without awareness, to reflect information contained in news reports or implied by interviewer questions.¹⁷⁸

A related line of research has focused on the effect of postwitnessing information on the inclination of a witness to be accurate versus informative.¹⁷⁹ Witnesses who are exposed to media reports that either promote the fidelity of eyewitness reports or denigrate it adopt a practice that comports.¹⁸⁰ Those who learn that eyewitness testimony is beneficial to criminal justice are inclined to be highly informative with a lower criterion for accuracy. By contrast, those who learn that eyewitness testimony is of questionable value are more circumspect and inclined toward accuracy.

175. Loftus, *supra* note 109; Loftus, *supra* note 88.

176. Lindsay & Johnson, *supra* note 138, at 349–58.

177. C. A. Elizabeth Luus & Gary L. Wells, *The Malleability of Eyewitness Confidence: Co-Witness and Perseverance Effects*, 79 J. Applied Psych. 714 (1994), <https://doi.org/10.1037//0021-9010.79.5.714>; Mitchell L. Eisen et al., “I Think He Had a Tattoo on His Neck”: How Co-Witness Discussions About a Perpetrator’s Description Can Affect Eyewitness Identification Decisions, 6 J. Applied Rsch. Memory & Cognition 274 (2017), <https://doi.org/10.1016/j.jarmac.2017.01.009>; Anders Granhag Pär et al., *Social Influence on Eyewitness Memory*, in *Forensic Psychology in Context: Nordic and International Approaches* 139 (2010), <https://doi.org/10.4324/9781315094038-8>.

178. Elizabeth F. Loftus & Guido Zanni, *Eyewitness Testimony: The Influence of the Wording of a Question*, 5 Bull. Psychonomic Soc. 86 (1975), <https://doi.org/10.3758/bf03336715>; Elizabeth F. Loftus, *Leading Questions and the Eyewitness Report*, 7 Cognitive Psych. 560 (1975), [https://doi.org/10.1016/0010-0285\(75\)90023-7](https://doi.org/10.1016/0010-0285(75)90023-7); Elizabeth F. Loftus, David G. Miller & Helen J. Burns, *Semantic Integration of Verbal Information into a Visual Memory*, 4 J. Exper. Psych.: Hum. Learning & Memory 19 (1978), <https://doi.org/10.1037/0278-7393.4.1.19>; Robin Blom & Kuo-Ting Huang, *Eyewitness Memory Contamination Through Misleading Questions by Reporters*, 42 News Rsch. J. 346 (2021), <https://doi.org/10.1177/07395329211030628>.

179. Morris Goldsmith, Asher Koriati & Amit Weinberg-Eliezer, *Strategic Regulation of Grain Size Memory Reporting*, 131 J. Exper. Psych.: Gen. 73 (2002), <https://doi.org/10.1037//0096-3445.131.1.73>; Nathan Weber & Neil Brewer, *Eyewitness Recall: Regulation of Grain Size and the Role of Confidence*, 14 J. Exper. Psych.: Applied 50 (2008), <https://doi.org/10.1037/1076-898x.14.1.50>.

180. Muhammad Mussaffa Butt, et. al, *Eyewitness Memory in the News Can Affect the Strategic Regulation of Memory Reporting*, 30 Memory 763 (2022), <https://doi.org/10.1080/09658211.2020.1846750>.

“Pristine” conditions for eyewitness interrogation are often sought in order to avoid these forms of social biasing,¹⁸¹ much in the way that restricted access to task-irrelevant information is intended to avoid bias in forensic analyses.¹⁸² Pristine is good and aspirational, of course, but in the real world it becomes very difficult to limit access to information and defuse explicit or implicit social pressures.¹⁸³

The long reach of modern social networks

The social networks of yore, which may have included school and work colleagues, or friends from the neighborhood, have become overshadowed today by vast online communities. Active participation in these communities frequently involves presenting yourself for validation, but the reflection is from a “trick mirror,”¹⁸⁴ which distorts who you are—often without your awareness—and makes powerful social demands on who you should be, what you should believe, and how you should behave.¹⁸⁵ The rewards for conformity are addictive.¹⁸⁶ Naturally, this form of social influence has profound implications for the validity of eyewitness identification, such that in the extreme, “some eyewitness identifications may not be governed by memory at all.”¹⁸⁷ Social pressures from online communities may cause a witness to question what they actually saw or, in the worst case, modify their memories without awareness. The “Rashomon effect” is the archetype of this phenomenon,¹⁸⁸ but we now see this happening on a massive scale in which scores of people collectively succumb to social distortions of reality and modify their criteria for judgment.¹⁸⁹

181. Wixted & Wells, *supra* note 25.

182. Nat’l Comm. Forensic Sci., Ensuring That Forensic Analysis Is Based Upon Task-Relevant Information (2015), <https://perma.cc/DT64-JLGM>.

183. *A National Survey of Eyewitness Identification*, *supra* note 11.

184. Jia Tolentino, Trick Mirror: Reflections on Self-Delusion (2019).

185. Jonathan Haidt, *The Anxious Generation: How the Great Rewiring of Childhood is Causing an Epidemic of Mental Illness* (2024).

186. *Id.*

187. Margaret Bull Kovera & Andrew J. Evelo, *Eyewitness Identification in its Social Context*, 10 J. Applied Rsch. Memory & Cognition 313 (2021), <https://doi.org/10.1016/j.jarmac.2021.04.003>.

188. *Rashomon Effect*, Wikipedia, <https://perma.cc/DYM4-8T86>.

189. Dustin P. Calvillo, Justin D. Harris & Whitney C. Hawkins, *Partisan Bias in False Memories for Misinformation About the 2021 U.S. Capitol Riot*, 31 Memory 137 (2023), <https://doi.org/10.1080/09658211.2022.2127771>; Seth Oranburg, *Social Media and Democracy After the Capitol Riot, or, a Cautionary Tale of the Giant Goldfish*, 73 Mercer L. Rev. 591 (2021); Brenna M. Davidson & Tetsuro Kobayashi, *The Effect of Message Modality on Memory for Political Disinformation: Lessons from the 2021 U.S. Capitol Riots*, 132 Computs. Hum. Behav. 107241 (2022), <https://doi.org/10.1016/j.chb.2022.107241>.

Large-scale social influences on perception and judgment are not new; the problem is simply exacerbated by online social media. One of the most extraordinary non-social media examples in recent history is the eyewitness report by thousands of devout Portuguese Catholics that on

In the end, there may be little that can be practically done beyond basic education on critical thinking to prevent large-scale social influences on memory and judgment. But law enforcement, the judiciary, juries, and witnesses themselves must be made aware of these growing risks in order to evaluate the probability that eyewitness testimony is accurate.

Evidence-Based Suspicion

The concern here is that a person may be placed as a suspect in a police lineup in the absence of any plausible evidence that the person had anything to do with the crime in question. There are no legal constraints that would prevent this practice. While there are few data that bear on the frequency with which it occurs, evidence indicates that it is not uncommon.¹⁹⁰ In fact, a recent survey of U.S. law enforcement agencies revealed that a significant fraction operate on the belief that they need no evidence at all, or perhaps little more than a hunch, before placing a person in a lineup.¹⁹¹ In 2020, the American Psychology-Law Society (AP-LS) charged six experts with developing modern policy and procedure recommendations for eyewitness identifications. Prominent among the recommendations in the AP-LS Report is the law enforcement use of “evidence-based suspicion” as a minimum criterion for lineup construction:

There should be evidence-based grounds to suspect that an individual is guilty of the specific crime being investigated before including that individual in an identification procedure and that evidence should be documented in writing prior to the lineup.¹⁹²

October 13, 1917, the midday “sun’s disc did not remain immobile. This was not the sparkling of a heavenly body, for it spun round on itself in a mad whirl when suddenly a clamor was heard from all the people. The sun, whirling, seemed to loosen itself from the firmament and advance threateningly upon the earth as if to crush us with its huge fiery weight.” John De Marchi, *The Immaculate Heart: The True Story of Our Lady of Fátima* (1952) (quote attributed to eyewitness Almeida Garrett). Astronomical records indicate that this “Miracle of the Sun” did not happen, but that is not what the witnesses remembered, reported, and incorporated into life decisions. See also Jeffrey S. Bennett, *When the Sun Danced: Myth, Miracles, and Modernity in Early Twentieth-Century Portugal* (2012).

190. John T. Wixted et al., *Estimating the Reliability of Eyewitness Identifications from Police Lineups*, 113 PNAS 304 (2016), <https://doi.org/10.1073/pnas.1516814112>; Bruce W. Behrman & Regina E. Richards, *Suspect/Foil Identification in Actual Crimes and in the Laboratory: A Reality Monitoring Analysis*, 29 Law & Hum. Behav. 279 (2005), <https://doi.org/10.1007/s10979-005-3617-y>.

191. Richard A. Wise, Martin A. Safer & Christina M. Maro, *What U.S. Law Enforcement Officers Know and Believe About Eyewitness Factors, Eyewitness Interviews and Identification Procedures*, 25 Applied Cognitive Psych. 488 (2011), <https://doi.org/10.1002/acp.1717>

192. Wells et al., *Policy and Procedure Recommendations* (2020), *supra* note 27.

This recommendation,¹⁹³ focused on law enforcement, is founded on the obvious but often overlooked fact that the simple act of placing someone in a lineup markedly increases the likelihood that they will be accused and prosecuted for a crime. Moreover, if no preexisting evidence ties that person to the crime in question, then the practice will naturally increase the probability that an innocent person is charged. There is also the worrisome question of what underlies a decision by law enforcement to include someone in a lineup in the absence of evidence for suspicion. Hunches of this sort may stem, in part, from the same prejudices that underlie biases in police traffic stops.¹⁹⁴ They may also arise from demographic information, including where a person lives relative to the crime scene, prior criminal history, or baseless anonymous tips, none of which rise to a level that could implicate that person, to the exclusion of others, as a legitimate suspect.

The resulting AP-LS recommendation that placement of a person in a lineup must be made only with “evidence-based suspicion” is thus a powerful strategy to avoid wrongful prosecution and conviction.¹⁹⁵ While this is a pre-emptive strategy that is necessarily under the control of the police, not the courts, knowledge of the practice used in any particular case can assist the court in evaluating the probability that the eyewitness testimony is correct.

Prior Exposure and Transference Effects

One feature of memory retrieval highlighted above under basic research (see section titled “Memory Storage” above) is the phenomenon of source memory failure and interference, in which prior knowledge from an “incorrect” source has been incorporated into a retrieved memory, which then biases memory-based decisions and impairs object recognition. This can happen in the context of eyewitness identification procedures employed by police if a witness has a prior opportunity to see a face that is part of a subsequent lineup. In the simplest case, the witness may be shown a mug shot of a potential suspect prior to the actual conduct of a lineup. The mug-shot face is then familiar to the witness at the time of a lineup that includes the face. But the witness may have lost all knowledge of why the face is familiar and naturally attributes it to their memory of the witnessed crime and the perpetrator. In simple terms, the critical memory

193. The recommendation was also made earlier by Gary Wells as the “reasonable suspicion criterion.” Gary L. Wells, *Eyewitness Identification: Systemic Reforms*, 2006 Wis. L. Rev. 615 (2006).

194. Emma Pierson et al., *A Large-Scale Analysis of Racial Disparities in Police Stops Across the United States*, 4 Nature Hum. Behav. 736 (2020), <https://doi.org/10.1038/s41562-020-0858-1>.

195. Failure to follow this recommendation would also seem to violate Fourth Amendment protection against unreasonable searches and seizures.

has been “contaminated,” without awareness, or it has been flat-out replaced by the memory of the mug-shot photo (a phenomenon termed transference).¹⁹⁶ The expected result, which has been well documented in applied eyewitness studies, is that identification decisions made under conditions of unconscious memory contamination or transference are frequently incorrect, and often made with inflated confidence.¹⁹⁷ Not surprisingly, the same effects can be elicited by showing a witness multiple lineups that contain the same suspect.¹⁹⁸

Of critical importance here is the need for the court to be aware of all prior identification procedures—a recommendation made by the NRC committee on eyewitness evidence,¹⁹⁹ and by authors of the recent AP-LS document on eyewitness policy and procedures²⁰⁰—as this information bears on the probability that the eyewitness testimony is correct. We note here that the 1968 Supreme Court, *Simmons v. United States*, fully appreciated the concern about memory contamination from photos and the implications for accuracy of testimony.²⁰¹

196. Jennifer E. Dysart et al., *Mug Shot Exposure Prior to Lineup Identification: Interference, Transference, and Commitment Effects*, 86 J. Applied Psych. 1280 (2001), <https://doi.org/10.1037//0021-9010.86.6.1280>.

197. Committee Report, Report to the Secretary of State for the Home Department of the Departmental Committee on Evidence of Identification in Criminal Cases (1976) [hereinafter Devlin Report]; Evan Brown, Kenneth A. Deffenbacher & William Sturgill, *Memory for Faces and the Circumstances of Encounter*, 62 J. Applied Psych. 311 (1977), <https://doi.org/10.1037/e666602011-038>; Kenneth A. Deffenbacher, Brian H. Bornstein & Steven D. Penrod, *Mugshot Exposure Effects: Retroactive Interference, Mugshot Commitment, Source Confusion, and Unconscious Transference*, 30 Law & Hum. Behav. 287 (2006), <https://doi.org/10.1007/s10979-006-9008-1>; Dysart et al., *supra* note 196; Graham Davies, John Shepherd & Hadyn Ellis, *Effects of Interpolated Mugshot Exposure on Accuracy of Eyewitness Identification*, 64 J. Applied Psych. 232 (1979), <https://doi.org/10.1037//0021-9010.64.2.232>.

198. Steblay et al., *supra* note 157.

199. Identifying the Culprit, *supra* note 12.

200. Wells et al., *Policy and Procedure Recommendations* (2020), *supra* note 27.

201. *Simmons v. United States*, 390 U.S. 377, 383–84 (1968) (“It must be recognized that improper employment of photographs by police may sometimes cause witnesses to err in identifying criminals. A witness may have obtained only a brief glimpse of a criminal, or may have seen him under poor conditions. Even if the police subsequently follow the most correct photographic identification procedures and show him the pictures of a number of individuals without indicating whom they suspect, there is some danger that the witness may make an incorrect identification. This danger will be increased if the police display to the witness only the picture of a single individual who generally resembles the person he saw, or if they show him the pictures of several persons among which the photograph of a single such individual recurs or is in some way emphasized. The chance of misidentification is also heightened if the police indicate to the witness that they have other evidence that one of the persons pictured committed the crime. Regardless of how the initial misidentification comes about, the witness thereafter is apt to retain in his memory the image of the photograph, rather than of the person actually seen, reducing the trustworthiness of subsequent lineup or courtroom identification.”).

Identification Procedures

Law enforcement can readily adopt procedures for how they present faces to an eyewitness. Few topics in applied eyewitness research have spawned as much scientific experimentation, debate, and impact on law enforcement practices, as how to best design these identification procedures. While the goal of these efforts has been to prospectively improve eyewitness performance to better serve the needs of criminal investigation and prosecution, knowledge of the identification procedures that were used in a given case can assist courts in retrospective evaluation of the accuracy of testimony.

Showups

In the simplest identification task, known as a “showup,” the suspect alone is presented to the witness, who responds yes or no. One common reason for conducting a showup is convenience and exigence: Shortly after arriving at the crime scene, the police apprehend a suspect fitting the witness’s description, who is then shown to the witness for confirmation. Despite the seeming benefits of this procedure, it is flawed, owing to suggestive qualities that may lower a witness’s criterion for an identification: (1) the witness’s knowledge that the suspect was apprehended nearby sets up an expectation, (2) the presentation of a sole suspect leads to an inference that the police must be quite certain, and (3) the impromptu context encourages the witness to be forthcoming with assistance. Under the circumstances, we should naturally expect that showups will increase the probability of correct identification (it’s much easier to identify your suitcase if it is the only one on the carousel), but they should also increase the probability of making an incorrect identification. Not surprisingly, applied eyewitness studies demonstrate that mistaken identifications are more common using the showup procedure than with “standard” six-person lineups,²⁰² particularly when innocent suspects resemble the witness’s description of the perpetrator.²⁰³

Confirmatory photo identification

Police will, on occasion, display a single photograph to a witness in an effort to confirm the identity of a person of interest. Police typically limit this method

202. A. Daniel Yarmey, Meagan J. Yarmey & Linda A. Yarmey, *Accuracy of Eyewitness Identifications in Showups and Lineups*, 20 *Law & Hum. Behav.* 459 (1996), <https://doi.org/10.1007/bf01498981>.

203. Nancy Steblay et al., *Eyewitness Accuracy Rates in Police Showup and Lineup Presentations: A Meta-Analytic Comparison*, 27 *Law & Hum. Behav.* 523, 523–40 (2003), <https://doi.org/10.1023/a:1025438223608>.

to situations in which the person is previously known to or acquainted with the witness. The goal is to clarify the legal identity (birth name/government name) of an individual who is well known to a witness, but only by a street name. In such examples, a witness may know (and may have known) the perpetrator for years but may be able to identify him only by a street name, such as “Diamond.” The police will typically use this procedure before making an arrest. As for mug books and showups, the procedure is highly suggestive, as it implicates the person in the photo, which may, in turn, unconsciously contaminate or replace the witness’s memory of the face of the perpetrator.²⁰⁴ The practical consequence of this procedure is spelled out above in the section titled “Prior Exposure and Transference Effects,” namely reduced identification accuracy in a subsequent lineup, and artifactually elevated confidence on the part of the witness.²⁰⁵ The existence of this practice highlights, once again, the need for courts to have knowledge of all prior identification procedures in a given case in order to evaluate their implications for eyewitness accuracy.²⁰⁶

Lineups: Simultaneous versus sequential procedures

To engage discriminative abilities and limit selection bias, multiperson lineups have been the standard since at least the nineteenth century. The people in the lineup commonly include the suspect and a set of people known as “fillers,” who are known to be innocent and (ideally) appear similar to the witness’s description of the perpetrator. In laboratory studies, “target absent” (innocent suspect) lineups are also used to enable a measure of the probability of identifying the known culprit relative to the probability of identifying an innocent person. In many studies conducted since the 1970s, this target present versus target absent comparison has been the standard approach to empirically evaluate the effects of estimator and system variables on eyewitness performance (many of which are cited in the foregoing sections).

In the 1980s, the discussion of factors that affect eyewitness performance broadened to include the nature of the multiperson lineup itself. Lindsay and Wells hypothesized that simultaneous viewing of lineup participants—the standard practice of the time—was a factor that prejudiced identifications.²⁰⁷ In particular, they posited that visual comparisons between faces (relative comparisons) could take precedence over the comparisons that really mattered—that is,

204. Wells et al., *Recommendations for Lineups and Photospreads*, *supra* note 27, at 457–70.

205. Deffenbacher et al., *Mugshot Exposure Effects*, *supra* note 197, at 287–307.

206. Identifying the Culprit, *supra* note 12; Devlin Report, *supra* note 197, at 3.

207. R.C. Lindsay & Gary L. Wells, *Improving Eyewitness Identifications from Lineups: Simultaneous Versus Sequential Lineup Presentation*, 70 J. Applied Psych. 556 (1985), <https://doi.org/10.1037//0021-9010.70.3.556>.

those between each lineup face and the visual memory of the perpetrator (absolute comparisons), leading to a selective increase in identifications of innocent suspects. To overcome this assumed problem, Lindsay and Wells proposed a novel lineup in which participants are viewed sequentially.

Early laboratory comparisons of performance on simultaneous versus sequential lineups appeared to demonstrate a sequential advantage, based on a higher ratio of correct to incorrect identifications.²⁰⁸ In lockstep with this science and mindful of wrongful convictions, many jurisdictions adopted the new sequential procedure.²⁰⁹ The problem with this outcome originates with the measure of performance—the diagnosticity ratio (DR)—which is affected both by the eyewitness’s ability to discriminate the perpetrator from an innocent suspect and by the criterion used to make an identification decision. It could have been true that sequential lineups also improve discriminability, which would be of real value, but it is impossible to know that using a single DR measure.

There is an established procedure for evaluating discriminability independently of the decision criterion. This procedure, known as analysis of receiver operating characteristics (ROC), is drawn from signal detection theory.²¹⁰ Figure 13 summarizes results from over 20 years of studies that have employed the ROC approach. Each dot represents data from a published study, identified at right. The x-axis plots a measure of the difference between discriminability measures obtained for the two lineup types, with positive values (SIM > SEQ) indicating a simultaneous advantage, and negative values (SEQ > SIM) indicating a sequential advantage. (Cohen’s *d* is a standard index of effect size.) Dot size reflects sample size (the number of people tested in the study); the y-axis orders studies as a function of sample size. Green dots indicate studies for which the discriminability difference for lineup types is statistically significant. Thus, contrary to initial claims for the superiority of the sequential lineup based on the DR measure, collective evidence supports a small but consistent improvement in discriminability for the simultaneous lineup relative to sequential.²¹¹ Extrapolation from the laboratory suggests that identifications made in jurisdictions that employ the sequential procedure may not be based on the best human discriminative abilities, which is a factor for consideration when assessing the probability that eyewitness testimony is accurate.

208. *Id.*

209. A National Survey of Eyewitness Identification Procedures, *supra* note 11.

210. Signal detection theory includes a framework for differentiating a person’s ability to discriminate the presence and absence of a stimulus from the criterion the person uses to make that decision. David M. Green & John A. Swets, *Signal Detection Theory and Psychophysics* (1966).

211. Travis M. Seale-Carlisle et al., *Designing Police Lineups to Maximize Memory Performance*, 25 J. Exper. Psych.: Applied 410 (2019), <https://doi.org/10.1037/xap0000222>; Laura Mickes & John T. Wixted, *Eyewitness Memory*, in 2294 *Oxford Handbook of Human Memory* (2022).

SEQ > SIM

SIM > SEQ

Anderson et al.

Mickes Flowe Wixted (Exp 1a)

Mickes Flowe Wixted (Exp 1b)

Meisters et al.

Goodsell position 5 (good)

Goodsell position 2 (good)

Goodsell position 5 (superior)

Goodsell position 2 (superior)

Horry et al.

Willing et al.

Seale-Carlisle et al. (Exp 5)

Seale-Carlisle et al. (Exp 1)

Seale-Carlisle Mickes

Rotello, Guggenmos, & Isbell

Carlson Carlson

Dobolyi Dodson (2 repetitions)

Dobolyi Dodson (4 repetitions)

Cohen's d

431

Other lineup procedures

One of the limitations of traditional lineup procedures is that eyewitnesses are asked to make two distinct contributions to the decision: (1) measurement of the similarity between memory and the face of a lineup participant, and (2) classification of the face as a match or nonmatch. An eyewitness identification manifests the classification decision, but the recipient of that decision has no direct access to the measurement upon which it was based, or to the criterion that the eyewitness used to evaluate the measurement, leaving the decision susceptible to unrecognized bias. This is notably distinct from man-made information processing devices (e.g., smartphone fingerprint and face detectors), which yield similarity measurements that are accessible and classifiable by an objective statistical procedure. A lineup procedure could, in principle, render a similarity measure but exclude the eyewitness from the classification—the latter performed instead by the lineup administrator using an objective statistical analysis. Recent laboratory studies demonstrate that this is a feasible, high-accuracy approach that could transform the utility of eyewitness lineups,²¹² but adoption by the criminal justice system must wait for evaluation in field studies with real casework, as we consider below.

In-court identifications

In criminal trials that are built on eyewitness testimony, a nearly ubiquitous feature is the theatrical event in which a witness is asked to identify the perpetrator in front of the jury. This practice has two consequences that adversely impact efforts by jurors to assess the probability that the witness testimony is correct.

First, the witness's level of confidence in the identification made at trial is generally highly inflated. In the Courtney case summarized above (see section titled “An Illustrative Case”), the victim-witness reported a confidence level of 60% upon initial identification. After identifying Courtney in court at the request of the prosecution, the witness expressed their confidence as: “I will never forget what he looks like.” The section titled “Timing of identification and conditions of confidence assessment,” above, delves into the specific reasons why this happens.

Second, the inflated confidence expressed during an in-court identification is not commonly recognized as such by a lay jury. It thus becomes a very powerful form of rhetoric that conveys authority on the part of the witness and stirs the

212. Sergei Gepshtein et al., *A Perceptual Scaling Approach to Eyewitness Identification*, 11 *Nature Commc'ns* 3380 (2020), <https://doi.org/10.1038/s41467-020-17194-5>; Sergei Gepshtein & Thomas D. Albright, *Thinking Outside the Lineup Box: Eyewitness Identification by Perceptual Scaling*, 10 *J. Applied Rsch. Memory & Cognition* 221 (2021), <https://doi.org/10.1016/j.jarmac.2021.05.001>.

emotions of the jurors.²¹³ All of which moves them toward a decision to convict.²¹⁴ As Supreme Court Justice William Brennan wrote, “[T]here is almost nothing more convincing than a live human being who takes the stand, points a finger at the defendant, and says ‘That’s the one!’”²¹⁵ This “convincing,” of course, is the primary reason why expert testimony or carefully worded jury instructions can be valuable, particularly when they explain the reasons *why* confidence is inflated.

In view of the potential for biasing the outcome, one might argue that in-court identifications should be eliminated altogether in criminal trials.²¹⁶ This proposal has gained some traction. The Massachusetts Supreme Judicial Court has ruled, for example, that in-court identifications should not normally be permitted unless they are the first identification, or if the out-of-court identification was suppressed.²¹⁷ The Connecticut Supreme Court has ruled to allow an in-court identification only if the witness knew the defendant before the crime, or if a prior lineup was proved to be nonsuggestive.²¹⁸ These rules have exceptions, as noted, and other courts conduct a burden-shifting approach toward in-court identifications; an in-court identification could also potentially be so unduly suggestive as to amount to a violation of *Manson v. Brathwaite*.²¹⁹ Research supports the view that in-court identifications should be handled with great caution, and that jurors must be made aware that they should not place such great weight on the in-court confidence of an eyewitness.

213. Maksymilian Del Mar, *The Confluence of Rhetoric and Emotion: How the History of Rhetoric Illuminates the Theoretical Importance of Emotion*, 36 Law & Literature 1 (2024), <https://doi.org/10.1080/1535685x.2022.2115651>; Aristotle, *The Rhetoric of Aristotle: A Translation* (Richard Claverhouse Jebb trans., 1909).

214. Brewer & Burke, *supra* note 173; Slane & Dodson, *supra* note 173.

215. *Watkins v. Sowders*, 449 U.S. 341, 353 (1981) (Brennan, J., dissenting) (citing Elizabeth F. Loftus, *Eyewitness Testimony* (1979)).

216. Brandon L. Garrett, *Eyewitnesses and Exclusion*, 65 Vand. L. Rev. 451 (2012).

217. See *Commonwealth v. Johnson*, 45 N.E.3d 83, 92–94 (Mass. 2016) (reasoning that “a subsequent in-court identification cannot be more reliable than the earlier out-of-court identification, given the inherent suggestiveness of in-court identifications and the passage of time”); *Commonwealth v. Crayton*, 21 N.E.3d 157, 169 (Mass. 2014) (“Where an eyewitness has not participated before trial in an identification procedure, we shall treat the in-court identification as an in-court showup, and shall admit it in evidence only where there is ‘good reason’ for its admission.”).

218. *State v. Dickson*, 141 A.3d 810, 817 (Conn. 2016) (“[I]n cases in which identity is an issue, in-court identifications that are not preceded by a successful identification in a nonsuggestive identification procedure implicate due process principles and, therefore, must be prescreened by the trial court.”). Note that this ruling has been challenged. See *Tatum v. Comm’r of Correction*, No. 20727, 2024 WL 3433265, at *3 (Conn. July 16, 2024).

219. See *State v. Hickman*, 330 P.3d 551, 568 (Or. 2014) (en banc) (“Courts considering the admissibility of first-time in-court identifications generally have placed the burden of seeking a prophylactic remedy on the defendant.”); *United States v. Domina*, 784 F.2d 1361, 1369 (9th Cir. 1986) (noting that district court’s denial of request for in-court lineup will only be overturned if in-court identification procedures were so suggestive and conducive to irreparable misidentification as to amount to denial of due process).

Filler Selection

Fillers are lineup participants known to be innocent, who serve as lures to challenge recognition memory. The choice of fillers has long been known to markedly influence eyewitness performance in laboratory studies.²²⁰ People recognize objects probabilistically based on the degree to which they elicit a memory signal corresponding to a particular target previously seen. It naturally follows that similar objects are more likely to elicit the same memory signal and thus have similar likelihoods of being recognized as the target. The similarity of fillers to the witness's description of the perpetrator is thus an important variable that affects both eyewitness discriminability and accuracy.

Lineups composed of fillers who are all of roughly the same degree of similarity to the witness's description are termed "fair." Conversely, lineups composed of fillers that possess differing degrees of similarity to the description are termed "unfair" or "biased." An unfair lineup, in which one filler is closer in similarity to the description of the perpetrator, reduces uncertainty by reducing the number of sensible choices, thus increasing the likelihood of correctly identifying the perpetrator, but also increasing the likelihood of misidentifying someone who looks like the perpetrator.²²¹

Despite the well-understood and potentially disastrous consequences of unfair lineups, there have been few attempts to systematize the law enforcement process of filler selection. Published guidelines for filler selection state that lineups should be constructed to ensure that "the suspect does not unduly stand out" and lineups should "avoid using fillers that so closely resemble the suspect that a person familiar with the suspect might find it difficult to distinguish the suspect from the fillers."²²² This guidance—fillers should be similar to the suspect but not too much so—is open to interpretation and is applied by different agents in different ways. It has not been effective.²²³

One emerging approach would use face similarity metrics to create lineups in which fillers are of known similarity to each other and to the witness's description of the perpetrator. A recent meta-analysis of laboratory studies adopting this approach revealed that decreasing the similarity between the suspect and fillers has the effect of increasing the probability that the suspect is identified,

220. Roy S. Malpass & Patricia G. Devine, *Measuring the Fairness of Eyewitness Identification Lineups*, in 81 *Evaluating Witness Evidence* (1983); Gary L. Wells, Michael R. Leippe & Thomas M. Ostrom, *Guidelines for Empirically Assessing the Fairness of a Lineup*, 3 *Law & Hum. Behav.* 285 (1979), <https://doi.org/10.1007/bf01039807>.

221. This effect is much like the consequence of a showup identification.

222. U.S. Dep't of Just., Nat'l Inst. of Just., *supra* note 27.

223. Carmen A. Lucas, Neil Brewer & Matthew A. Palmer, *Eyewitness Identification: The Complex Issue of Suspect-Filler Similarity*, 27(2) *Psych., Pub. Pol'y, & L.* 151 (2021).

regardless of whether that identification is correct.²²⁴ The case of Uriah Courtney, highlighted above, illustrates the tragic consequences of this practice in the real world: Courtney's conviction was based on two witness identifications using a lineup in which Courtney stood out from many of the fillers as an exemplar of the witness descriptions.²²⁵ By contrast, increasing the similarity between suspect and fillers increases the probability of filler identification or a lineup rejection, meaning that fewer suspects are identified overall. A more recent analysis suggests that this is a consequence of reduced discriminability between perpetrator and innocent suspect.²²⁶

A recent laboratory study adds a new wrinkle to this picture and a promising target for future investigation: Testing the counterintuitive predictions of a signal detection model, the investigators found that "choosing fillers who match the description of the perpetrator but who are otherwise dissimilar to the suspect's appearance yielded fair lineups and enhanced eyewitness identification performance."²²⁷ To appreciate this conclusion, suppose that the witness's verbal description of the perpetrator included red hair and a goatee. A suspect with those features is apprehended, and all lineup fillers are selected to have the same diagnostic features (i.e., the lineup is fair). The perpetrator may have other features not mentioned by the witness, such as blue eyes. Fillers may share the blue eyes feature with the perpetrator by chance, increasing the probability that the witness will misidentify them. Thus, the lineup is more likely to yield a correct outcome if fillers are chosen without regard to features that were not part of the witness's description of the perpetrator.

Recent evidence points to another variable associated with filler–suspect similarity that has a somewhat counterintuitive effect on eyewitness performance: the size of the face database from which fillers are selected. One might expect that having a large library of faces to draw from, such as computer databases of driver's license photos, would make it possible to find filler faces that are more suitable for a lineup and may facilitate eyewitness performance.²²⁸ Contrary to this intuition, selection of fillers from very large databases is commonly associated with greater similarity between the suspect and lineup fillers, which has the consequence of increasing the frequency of filler picks by eyewitnesses, thus reducing accuracy.²²⁹

224. Ryan J. Fitzgerald et al., *The Effect of Suspect-Filler Similarity on Eyewitness Identification Decisions: A Meta-Analysis*, 19 Psych., Pub. Pol'y, & L. 151 (2013), <https://doi.org/10.1037/a0030618>.

225. Albright, *Why Eyewitnesses Fail*, *supra* note 14, at 7758–64.

226. Fitzgerald et al., *supra* note 224, at 151.

227. Melissa F. Colloff et al., *Optimizing the Selection of Fillers in Police Lineups*, 118 PNAS e2017292118 (2021), <https://doi.org/10.1073/pnas.2017292118>.

228. Fitzgerald et al., *supra* note 224, at 151.

229. Amanda N. Bergold & Paul Heaton, *Does Filler Database Size Influence Identification Accuracy?*, 42 Law & Hum. Behav. 227 (2018), <https://doi.org/10.1037/lhb0000289>.

With a movement toward automated systems for filler selection, their ease of use, and their adoption by the police,²³⁰ database size is a critical variable.

There are many more details to be worked out in this space of filler–suspect similarity, but on the whole, it is clear that filler selection matters greatly to eyewitness performance, and thus the approach that was taken in a given case may offer insight into the probability that the eyewitness testimony is correct.

Corroborating Variables: Machine-Based Facial Recognition

There are, of course, many other variables associated with eyewitness events that fall in neither the estimator nor system category. For police investigations and criminal trials involving an eyewitness, the most useful of these variables are sources of forensic evidence that bear on, and potentially corroborate, eyewitness identification. The forms of evidence in this category are numerous, including, but not limited to, fingerprints, tool marks, chemical/biological residues, and DNA. The utility of these forms of evidence for the courts is treated elsewhere.²³¹

One specific and very recent form of corroborating evidence deserves mention here: automated facial recognition. New artificial intelligence (AI) technology for machine-based person recognition derived from photographs or video collected in the real world, combined with the dramatic proliferation of live cameras in public spaces, has enabled law enforcement to harvest identification evidence that may complement traditional eyewitness testimony. This is a rapidly evolving and largely unregulated field, which has considerable relevance to judicial decisions about the accuracy of eyewitness testimony from real people. We refer readers to a report from the National Academies of Sciences, Engineering, and Medicine, which details the technology and its accuracy, utility, potential applications, and privacy concerns.²³²

Applied Eyewitness Studies in the Field

The results of applied studies conducted in the laboratory have revealed a large number of variables that affect the ability of witnesses to discriminate between the perpetrator and innocent suspects, as well as the overall accuracy of

230. A National Survey of Eyewitness Identification Procedures, *supra* note 11.

231. See, e.g., Valena E. Beety, Jane Campbell Moriarty & Andrea L. Roth, *Reference Guide on Forensic Feature Comparison Evidence*, in this manual.

232. Nat'l Acads. Scis., Eng'g & Med., *Facial Recognition Technology: Current Capabilities, Future Prospects, and Governance* (2024), <https://doi.org/10.17226/27397>.

identifications made. These findings both encourage this empirical strategy and hold promise for reform, but they have often been criticized for lack of ecological validity.²³³ Specifically, the laboratory crime is a recognized pretense that cannot be expected to elicit critical states of the observer, such as high values of stress, fear, pain, and loss of attentional focus. The ultimate test of the utility of laboratory discoveries is thus successful translation to real criminal casework as evaluated by field studies.²³⁴

There are three constraints associated with achieving this translational goal. First, it requires access to a high-functioning police department with a willingness to test scientific hypotheses. Second, ground truth is unknown in real cases, which means that experimental efforts must rely on independent corroboration of truth (“extrinsic evidence”) as a means to validate the effects of the variables tested. That reliable corroboration is often difficult to come by. Third, the situational context in the field is difficult to control with the level of precision that is common in a laboratory. Real-world things sometimes happen unexpectedly, or there may simply be no record of what did happen.

Operating under these constraints, a small number of studies have tested the effects of estimator and system variables in the field. Results generally support findings from laboratory studies of effects of retention interval, cross-race effect, weapon focus, and lineup type,²³⁵ but many factors have been poorly

233. Yoojin Chae, *Application of Laboratory Research on Eyewitness Testimony*, 10 J. Forensic Psych. Prac. 252 (2010), <https://doi.org/10.1080/15228930903550608>; Graham F. Wagstaff et al., *Can Laboratory Findings on Eyewitness Testimony Be Generalized to the Real World? An Archival Analysis of the Influence of Violence, Weapon Presence, and Age on Eyewitness Accuracy*, 137 J. Psych. 17 (2003), <https://doi.org/10.1080/00223980309600596>.

234. Daniel L. Schacter et al., *Policy Forum: Studying Eyewitness Investigations in the Field*, 32 Law & Hum. Behav. 3 (2008), <https://doi.org/10.1007/s10979-007-9093-9>.

235. John C. Yuille & Judith L. Cutshall, *A Case Study of Eyewitness Memory of a Crime*, 71 J. Applied Psych. 291 (1986), <https://doi.org/10.1037//0021-9010.71.2.291>; Yarmey et al., *supra* note 202, at 459–77; Bruce W. Behrman & Sherrie L. Davey, *Eyewitness Identification in Actual Criminal Cases: An Archival Analysis*, 25 Law & Hum. Behav. 475 (2001), <https://doi.org/10.1023/a:1012840831846>; Amy Klobuchar, Nancy K. Mehrkens Steblay & Hilary Lindell Caligiuri, *Improving Eyewitness Identifications: Hennepin County's Blind Sequential Lineup Pilot Project*, 4 Cardozo Pub. L., Pol'y & Ethics J. 381 (2006); Daniel L. Schacter et al., *Policy Forum: Studying Eyewitness Investigations in the Field*, 32 L. & Hum. Behav. 3 (2008), <https://doi.org/10.1007/s10979-007-9093-9>; Nancy K. Steblay, *What We Know Now: The Evanston Illinois Field Lineups*, 35 Law & Hum. Behav. 1 (2011), <https://doi.org/10.1007/s10979-009-9207-7>; Gary L. Wells, Nancy K. Steblay & Jennifer E. Dysart, *A Test of the Simultaneous vs. Sequential Lineup Methods: An Initial Report of the AJS National Eyewitness Identification Field Studies*, Am. Judicature Soc. (2011), <https://perma.cc/42F5-LVL5>; Gary L. Wells, Nancy K. Steblay & Jennifer E. Dysart, *Double-Blind Photo Lineups Using Actual Eyewitnesses: An Experimental Test of a Sequential Versus Simultaneous Lineup Procedure*, 39 Law & Hum. Behav. 1 (2015), <https://doi.org/10.1037/lhb0000096>; Karen L. Amendola & John T. Wixted, *Comparing the Diagnostic Accuracy of Suspect Identifications Made by Actual Eyewitnesses from Simultaneous and Sequential Lineups in a Randomized Field Trial*, 11 J. Exper. Crim. 263 (2015), <https://doi.org/10.1007/s11292-014-9219-2>; Wixted et al., *supra* note 190.

controlled. Moreover, in real police investigations it can be hard to assess ground truth, and cognitive biases can play a complicating role in perceptions of evidence strength. Much more needs to be done in this difficult arena before changes are made to policy and practice. It seems possible that the laboratory versus field relationship will break down where the real emotion of criminal witnessing comes into play.

Effects of Eyewitness Testimony on Juries

The identification decision made by an eyewitness is only the first stage in which human factors may drive a criminal prosecution off track. The second stage is in the communication of eyewitness testimony to the trier of fact. Recent scientific studies have explored the role that communication plays in inferring the accuracy of eyewitness testimony. Specifically, the witness is viewed as the transmitter of information, and the jury is the intended receiver of that information. While much science focuses on improving the accuracy of testimony and the ability to predict that accuracy, these efforts will be in vain if the information is not correctly received by the trier of fact.

Laypeople may have a range of misconceptions regarding how eyewitness memory functions. Further, there is a traditional practice of allowing eyewitnesses to identify a person in the courtroom (see section titled “In-court identifications” above). As Justice Sonia Sotomayor explained in dissent in *Perry v. New Hampshire*: “At trial, an eyewitness’ artificially inflated confidence in an identification’s accuracy complicates the jury’s task of assessing witness credibility and reliability.”²³⁶ Research has confirmed that these in-court identifications powerfully influence jurors, and in addition, that the confidence of an eyewitness in court has a particularly outsized influence on lay jurors.²³⁷ An eyewitness may appear to be highly confident in court, even if the witness was highly uncertain during the earlier eyewitness identification procedure conducted by police out of court—the phenomenon known as confidence inflation (see above section titled “Timing of identification and conditions of confidence assessment”).²³⁸ Indeed, confidence can be inflated by biasing information, including feedback that eyewitnesses receive after an identification procedure and in preparation for testimony at trial.²³⁹

236. 565 U.S. 228, 252 (2012), (Sotomayor, J., dissenting).

237. See generally, Garrett et al., *supra* note 33 (detailing findings of mock juror survey regarding weight given to confidence of eyewitnesses).

238. See Wells et al., *Policy and Procedure Recommendations* (2020), *supra* note 27, at 21.

239. Carl Martin Allwood, Jens Knutson & Pär Anders Granhag, *Eyewitnesses Under Influence: How Feedback Affects the Realism in Confidence Judgements*, 12 *Psych. Crime & L.* 25, 36 (2006), <https://doi.org/10.1080/10683160512331316316>.

As noted below, some jurisdictions and courts have adopted jury instructions and other mechanisms to improve communication between an eyewitness or their proxy (transmitter) and a jury (receiver), to better educate jurors regarding the limitations of eyewitness testimony, including by limiting the use of in-court identifications. There is, however, evidence that several of these more recent sets of jury instructions are not particularly effective at improving jurors' ability to discriminate between more or less reliable eyewitness identifications,²⁴⁰ perhaps because they do not sufficiently address the weight that jurors place on the courtroom confidence of the eyewitness, and—as highlighted above—the rhetorical impact of an in-court identification. Indeed, recent evidence shows that more pointed instructions regarding in-court confidence, and the reasons why there are dangers of relying on it, may be more effective.²⁴¹ Expert witnesses, of course, may address these issues head-on, and with a richer description of the relevant scientific research.

Sources of Expertise on Matters of Eyewitness Science

The value and utility of eyewitness evidence lie in both the content of testimony (who did what to whom) and the probability that the testimony is correct. As we have emphasized throughout this text, there are both general and specific features of the process of perceiving criminal events, identifying perpetrators, and testifying to experience, that bear on the probability that testimony is accurate. Knowledge and understanding of these features and their predictive relationship to accuracy are often beyond the ken of jurors, judges, and witnesses proffering testimony. In addition to numerous court rulings on this topic (particularly *Manson v. Brathwaite*, but also more recent federal and state court decisions that digest and rely on scientific research in the area), there are many sources of clarifying information that can help courts decide about admissibility of eyewitness testimony and assess its probative value.

These include written documents, such as (but not limited to) the present reference guide, the 2014 NRC report on eyewitness identification,²⁴² the 2019

240. Jones & Bergold, *supra* note 33, at 433–55; Berman, *supra* note 33; Marlee Kind Dillon et al., *Henderson Instructions: Do They Enhance Evidence Evaluation*, 17 J. Forensic Psych. Rsch. & Prac. 1 (2017); Garrett et al., *supra* note 33; Jones et al., *supra* note 33; Laub et al., *supra* note 33; Papailiou et al., *supra* note 33.

241. Brandon L. Garrett et al., *Sensitizing Jurors Regarding Eyewitness Testimony*, 12 J. Applied Rsch., Memory & Cognition 1, 141–57 (2022), <https://doi.org/10.1037/mac0000035>.

242. Identifying the Culprit, *supra* note 12.

report of the Third Circuit task force on eyewitness identifications,²⁴³ and the eyewitness policy and procedural recommendations of the AP-LS.²⁴⁴ The underpinning of these documents is a very large peer-reviewed scientific literature from basic and applied fields of psychology and/or cognitive and neural sciences, with particular focus on vision, memory, attention, signal detection, predictive coding, and decision-making (much of which is cited herein). In addition to these written guides, which often themselves require clarification and interpretation, decisions about admissibility of eyewitness evidence and assessment of its weight (probability that the evidence is correct) can be informed by experts with demonstrated training and proficiency in the relevant fields of science. General standards for admissibility of such experts are, of course, defined by the Supreme Court's *Daubert*²⁴⁵ ruling and Rule 702 of the Federal Rules of Evidence, which emphasize both qualifications of the expert and the validity and reproducibility of the underlying science reported by the expert.²⁴⁶

On the specific topic of eyewitness evidence, expert qualifications and experience should include a published record of scholarly empirical work on phenomenology, function, and underlying mechanisms of visual perception and recognition memory, applied studies of the effects of contextual variables (viewing conditions and cognitive state of the observer) on eyewitness discriminability and accuracy, and probabilistic modeling of evidence-based behavioral choice under conditions of uncertainty. Experts with these qualifications and knowledge would commonly hold tenured or tenure-track positions at accredited universities or in nonprofit research institutes, in academic disciplines of psychology and/or cognitive and neural sciences, and statistics. Such experts can draw clear and useful (and sometimes precisely quantitative) inferences about (but not limited to) the following: (1) the effects of contextual variables on initial eyewitness collection of information about criminal events, such as atmospheric conditions (e.g., rain, fog), viewing duration, viewing distance, illuminant intensity, luminance contrast, viewing angle, image “crowding,” distracting features that hijack attention, expectations, stress, and emotional states; (2) the degradation and contamination of memories of witnessed events with the passage of time; (3) the influence of particular behavioral strategies (including introduction of uncertainty and bias) for eliciting eyewitness recognition memory (e.g., lineup type, filler selection); (4) the social influence of other parties (e.g., law enforcement officers, attorneys, news media, friends, family) on choice and confidence in

243. *Report of the United States Court of Appeals for the Third Circuit Task Force on Eyewitness Identifications*, *supra* note 29.

244. Wells et al., *Policy and Procedure Recommendations* (2020), *supra* note 27, at 3–36.

245. *Daubert v. Merrell Dow Pharms., Inc.*, 509 U.S. 579 (1993).

246. Thomas D. Albright, *A Scientist's Take on Scientific Evidence in the Courtroom*, 120 PNAS e2301839120 (2023), <https://doi.org/10.1073/pnas.2301839120>.

proffered testimony; (5) the utility of self-report measures (e.g., confidence) that may be presented as indicia of accuracy; and (6) the consequences of poor communication between eyewitness (and counsel) and trier of fact.

The Legal Framework for Eyewitness Evidence

The potential inaccuracy of eyewitness evidence has long been well known to judges. As the U.S. Supreme Court put it in its 1967 ruling in *United States v. Wade*, “The vagaries of eyewitness identification are well-known; the annals of criminal law are rife with instances of mistaken identification.”²⁴⁷ The traditional constitutional criminal procedure approach to such evidence has involved several related protections, including the right to counsel at an in-person lineup, as well as a due process rule requiring a deferential review of the validity of an unduly suggestive lineup. In addition, courts have conducted pretrial hearings to examine the validity of eyewitness evidence, and offered fairly basic and brief jury instructions to explain considerations relevant to eyewitness evidence to jurors. Those rules were developed in the late 1960s and 1970s, and in federal courts, the basic framework has remained largely unchanged ever since.

In the nearly five decades since, a substantial body of visual perception and eyewitness memory science, as described above, has transformed our understanding of eyewitness evidence. This research has affected police identification practices, including those of federal law enforcement agencies. In federal court, these developments have informed, for example, a move toward receptivity to expert testimony regarding the science of eyewitness vision and memory, where it would assist the jury and satisfy other requirements regarding expert testimony,²⁴⁸ with other courts preferring the use of jury instructions.²⁴⁹ More recently, a noteworthy Third Circuit Task Force on Eyewitness Evidence examined a range of recommendations regarding how to consider eyewitness evidence.²⁵⁰

247. *United States v. Wade*, 388 U.S. 218, 228 (1967).

248. See generally, *United States v. Rodriguez-Felix*, 450 F.3d 1117, 1124 (10th Cir. 2006), (summarizing “there has been a trend in recent years to allow such testimony” and “subject to the trial court’s careful supervision, properly conceived expert testimony may be admissible to challenge or support eyewitness evidence” when it would assist the jury); *United States v. Harris*, 995 F.2d 532, 534 (4th Cir. 1993) (summarizing “[u]ntil fairly recently, most, if not all, courts excluded expert psychological testimony on the validity of eyewitness identification” but describing how courts came to recognize trial court discretion to allow such testimony).

249. See, e.g., *United States v. Jones*, 689 F.3d 12, 20 (1st Cir. 2012) (finding it “within the district court’s province to provide this information through instructions rather than through dueling experts”).

250. *Report of the Third Circuit Task Force on Eyewitness Identifications*, *supra* note 27.

State judges have adopted new frameworks for considering admissibility that incorporate new scientific developments: through interpretation of state constitutional provisions, through evidence law, by tasking state commissions with revising jury instructions, and through the judges' supervisory power. A series of courts have concluded, for example, "as a matter of state constitutional law, that it is appropriate to modify [due process standards] to conform to recent developments in social science and the law."²⁵¹ Courts increasingly recognize that vision is limited by uncertainty and bias, where "research further reveals that an array of variables can affect and dilute memory and lead to misidentifications."²⁵² State courts have adopted new jury instructions for eyewitness evidence, new standards for in-court identifications, and new admissibility standards, as well as requiring that pretrial reliability hearings (validity hearings) be conducted.²⁵³ As in federal courts, most state courts now permit, within the discretion of the trial judge, the appointment of expert witnesses on eyewitness perception and memory and may find it to be an error to exclude an eyewitness expert that would play an important role in educating the jury.²⁵⁴ Further, state lawmakers have been quite active in this area, enacting a series of statutes that regulate eyewitness identification procedures, and also, to a lesser extent, how judges review such evidence when procedures do not comply with these statutes.²⁵⁵

The sections that follow describe: (1) the U.S. Supreme Court's constitutional rulings regarding eyewitness evidence; (2) the rulings by federal courts of appeals and lower federal courts; (3) state court rulings, including those creating new frameworks for reviewing eyewitness evidence, and those regarding expert evidence and jury instructions; and (4) state statutes regarding eyewitness evidence.

The U.S. Supreme Court

The U.S. Supreme Court first examined the constitutional criminal procedure implications of eyewitness identifications in a trilogy of cases decided in 1967. In *Gilbert v. California*²⁵⁶ and *United States v. Wade*,²⁵⁷ the Court held that, once indicted, a person has a right under the Sixth Amendment to have a defense lawyer present at an in-person live lineup.²⁵⁸ Further, in *Wade*, the Court held that a failure to provide defense counsel at a postindictment lineup does not

251. *State v. Harris*, 191 A.3d 119, 134 (Conn. 2018).

252. *State v. Henderson*, 27 A.3d 872, 247 (N.J. 2011).

253. See Thomas D. Albright & Brandon L. Garrett, *The Law and Science of Eyewitness Evidence*, 102 Boston U. L. Rev. 511, 591 & Appendix B (2022) (surveying state court rulings).

254. See *id.* at 593 & Appendix C (surveying leading state cases).

255. See *id.* at 580 & Appendix A (surveying state statutes).

256. *Gilbert v. California*, 388 U.S. 263 (1967).

257. *United States v. Wade*, 388 U.S. 218, 228 (1967).

258. *Id.* at 236–38; *Gilbert*, 388 U.S. at 272.

Reference Guide on Eyewitness Identification

result in suppression of the evidence if the witness had an “independent source” for the identification, based on factors including the opportunity to observe the culprit at the crime scene, any discrepancies in witness descriptions, any prior identifications or failure to identify the person, and the lapse of time between the act and the lineup.²⁵⁹ The Court later held, in its 1973 ruling in *United States v. Ash*, that right to counsel does not extend to photo array procedures, which, today, police use far more than live or in-person lineups.²⁶⁰

In *Stovall v. Denno*,²⁶¹ the third case in the 1967 trilogy, the Supreme Court held that the Due Process Clause also regulates eyewitness identification procedures, stating that certain procedures may be “so unnecessarily suggestive” that the identification evidence must be suppressed.²⁶² However, the Court rejected any per se rule against the use of showup identifications and held that, based on a review of the totality of the circumstances, such identifications could be admissible despite the risks of suggestion.²⁶³ Further, in its 1968 ruling in *Simmons v. United States*,²⁶⁴ the Court emphasized the overall reliability of the eyewitness’s identification in finding no grounds for suppressing the identification under the due process theory announced in *Stovall*.²⁶⁵

In 1972, in *Neil v. Biggers*,²⁶⁶ the Court found, in a case involving a trial that occurred pre-*Stovall*, that even if an identification was conducted in an unnecessarily suggestive manner, a court should consider five factors in examining whether there was a “likelihood of misidentification” and a due process remedy warranted:

the opportunity of the witness to view the criminal at the time of the crime, the witness’ degree of attention, the accuracy of the witness’ prior description of the criminal, the level of certainty demonstrated by the witness at the confrontation, and the length of time between the crime and the confrontation.²⁶⁷

In 1977, in *Manson v. Brathwaite*, the Supreme Court adopted this *Biggers* test for all cases, as a general due process rule for consideration of eyewitness identification evidence potentially affected by undue law enforcement suggestion.²⁶⁸ The Court emphasized that “reliability is the linchpin in determining the admissibility of identification testimony.”²⁶⁹ Although the Court’s due process rule

259. *Wade*, 388 U.S. at 241.

260. *United States v. Ash*, 413 U.S. 300, 321 (1973).

261. *Stovall v. Denno*, 388 U.S. 293 (1967).

262. *Id.* at 301–02.

263. *Id.*

264. *Simmons v. United States*, 390 U.S. 377 (1968).

265. *Id.* at 378, 384–86.

266. *Neil v. Biggers*, 409 U.S. 188 (1972).

267. *Id.* at 199–200.

268. *Manson v. Brathwaite*, 432 U.S. 98, 114 (1977) (holding *Biggers* factors should be used to assess reliability).

269. *Id.*

asks whether police used suggestive identification procedures, any such suggestiveness can be excused based on a review of the five “reliability” factors set out in *Biggers*.²⁷⁰ The Court did not assign any particular weight to these factors.²⁷¹

In the years since the ruling, scientific research has advanced considerably, and in ways that might inform how a court applies the test. One central concern that scientists have raised concerning this test is that it could permit judges to excuse suggestive eyewitness identification procedures, based on an assumption that other indicia of reliability might counteract the influence of police suggestion.²⁷² The NRC report emphasized:

[T]he test treats factors such as the confidence of a witness as independent markers of reliability when, in fact, it is now well established that confidence judgments may vary over time and can be powerfully swayed by many factors.²⁷³

However, in applying the test, the *Biggers* factors need not be considered as “independent,” and judges have discretion regarding what weight to attach to them. Thus, the application of the *Manson* test could be informed by more recent scientific research.

Subsequent rulings by the Supreme Court have not altered the test adopted in *Manson*, but they have added further discretion and also limitations to the inquiry. In *Watkins v. Sowders*, the Supreme Court held that an evidentiary hearing regarding identification testimony, outside the presence of the jury, need not be conducted in all cases, but remains within the discretion of the trial judge.²⁷⁴ The Court noted that it may often be advisable to do so outside the presence of the jury but that in some situations it would not be needed, and cross-examination of the eyewitness can preserve a defendant’s due process rights.²⁷⁵

Most recently, the Supreme Court held in *Perry v. New Hampshire* that when eyewitness misidentifications are not due to intentional police action, the Due Process Clause does not apply.²⁷⁶ In that case, the suspect was detained near the

270. *Id.*

271. *Id.* at 116 (stating only that *Biggers* factors are “for the jury to weigh”).

272. Wells & Quinlivan, *supra* note 20, at 16 (condemning federal courts’ application of *Manson* test to find that identifications were reliable even when surrounding procedures were “highly suggestive”).

273. See Identifying the Culprit, *supra* note 12.

274. *Watkins v. Sowders*, 449 U.S. 341, 349 (1981).

275. *Id.* For examples of lower federal courts emphasizing that cross-examination sufficiently addressed reliability concerns, see, e.g., *United States v. Maloney*, 513 F.3d 350, 356 (3d Cir. 2008); *United States v. Saint Louis*, 889 F.3d 145, 154–55 (4th Cir. 2018); *Amador v. Quarterman*, 458 F.3d 397, 415 (5th Cir. 2006); *United States v. Stokes*, 631 F.3d 802, 807 (6th Cir. 2011).

276. *Perry v. New Hampshire*, 565 U.S. 228, 245 (2012) (“The fallibility of eyewitness evidence does not, without the taint of improper state conduct, warrant a due process rule requiring a trial court to screen such evidence for reliability before allowing the jury to assess its creditworthiness.”).

crime scene and the eyewitness looked out of her apartment, saw him there, and made an identification.²⁷⁷ The majority did note that they “do not doubt either the importance or the fallibility of eyewitness identifications” but maintain that state evidence law and safeguards such as expert testimony and jury instructions should be relied on to ensure accurate presentation of evidence.²⁷⁸

Lower Federal Court Rulings

In 1972, in *United States v. Telfaire*, the Court of Appeals for the District of Columbia Circuit crafted particularly influential jury instructions regarding eyewitness evidence.²⁷⁹ The instructions explain that jurors “must consider the credibility of each identification witness in the same way as any other witness, consider whether he is truthful, and consider whether he had the capacity and opportunity to make a reliable observation on the matter covered in his testimony.” The instructions ask jurors to: examine the “circumstances under which the identification was made”; “scrutinize the identification with great care”; “consider the length of time that lapsed between the occurrence of the crime and the next opportunity of the witness to see the defendant”; and consider any instances in which the witness failed to identify the defendant.

The instructions are fairly brief and generically touch on certain factors that can be relevant to assessing the reliability of eyewitness evidence. The *Telfaire* ruling was a step forward at the time, encouraging judges to provide additional context regarding eyewitness evidence. To be sure, the ruling predates the body of modern research on the subject, and as we will describe, courts have more recently offered instructions explicitly grounded in scientific research. While most federal courts use the *Telfaire* instructions,²⁸⁰ other federal courts adopt a “flexible approach” that provides district courts with discretion whether to use eyewitness jury instructions should they conclude, based on “strong reliability” using the *Biggers/Manson* factors, that no such instruction is necessary.²⁸¹ Thus, those federal courts rely on the factors set out in *Manson v. Brathwaite* in deciding whether to give the jury the *Telfaire* jury instructions.²⁸² Some federal courts

277. *Id.*

278. *Id.*

279. *United States v. Telfaire*, 469 F.2d 552 (D.C. Cir. 1972) (per curiam).

280. *United States v. Anderson*, 739 F.2d 1254, 1258 (7th Cir. 1984); *see also* *United States v. Greene*, 591 F.2d 471, 475 (8th Cir. 1979).

281. *See, e.g., United States v. Luis*, 835 F.2d 37, 41 (2d Cir. 1987) (“We believe this flexible approach remains the better course because it avoids imposing rigid requirements on trial courts under the threat that failure to give the requested charge will later be grounds for automatic reversal.”).

282. *Telfaire*, 469 F.2d at 558–59 (providing model jury special instructions on identification).

have also modestly supplemented the *Telfaire* instructions to include additional factors, such as whether it was a cross-racial identification.²⁸³

On the question of admissibility of expert evidence regarding eyewitness memory, federal courts, like state courts, have in recent years become more receptive to the use of such experts.²⁸⁴ Thus, in 1999, the Eighth Circuit explained it was “especially hesitant” to find it an abuse of discretion not to admit eyewitness expert testimony unless the case rests “exclusively on uncorroborated eyewitness testimony.”²⁸⁵ Other courts, however, emphasized the usefulness of eyewitness expert testimony and found it at times an abuse of discretion not to permit the defense such an expert.²⁸⁶ Pre-*Daubert*, the Eleventh Circuit was the only federal court of appeals that had adopted a per se rule that eyewitness expert evidence would *not* be permitted.²⁸⁷ The Eleventh Circuit still maintains that approach, while every other court of appeals now permits such experts, subject to the discretion of the district judge.²⁸⁸

More recent federal rulings regarding experts on eyewitness evidence have relied more heavily on modern scientific research to explain the need to provide such research to the jury. For example, the Second Circuit, in its 2021 ruling in *United States v. Nolan*, found defense counsel per se ineffective for failing to call an expert witness on eyewitness evidence. The panel explained: “*Strickland* ordinarily does not require defense counsel to call any particular witness,” but given

283. Comm. on Model Crim. Jury Instructions for the Third Circuit, Model Criminal Jury Instructions § 4.15 (2017).

284. For an overview, see Lauren Tallent, *Through the Lens of Federal Evidence Rule 403: An Examination of Eyewitness Identification Expert Testimony Admissibility in the Federal Circuit Courts*, 68 Wash. & Lee L. Rev. 765, 777–78 (2011).

285. See *United States v. Villiard*, 186 F.3d 893, 895 (8th Cir. 1999).

286. See, e.g., *United States v. Rodriguez–Felix*, 450 F.3d 1117, 1125 (10th Cir. 2006) (“[A]n expert’s testimony describing how certain factors, falling outside a typical juror’s experience, may affect an eyewitness’s identification is the very type of scientific knowledge to which *Daubert*’s relevance prong is addressed.”); *United States v. Stevens*, 935 F.2d 1380, 1400 (3d Cir. 1991) (reversing where misidentification was the key issue at trial and expert’s proposed testimony was outside the realm of typical juror knowledge); *United States v. Moore*, 786 F.2d 1308, 1313 (5th Cir. 1986) (“In some cases casual eyewitness testimony may make the entire difference between a finding of guilt or innocence. In such a case expert eyewitness identification testimony may be critical.”); *United States v. Smith*, 736 F.2d 1103, 1106 (6th Cir. 1984) (recognizing that expert testimony on eyewitness identification “might have refuted [jurors’] otherwise common assumptions about the reliability of eyewitness identification”); see also *Bell v. Miller*, 500 F.3d 149, 155–56 (2d Cir. 2007); *United States v. Brownlee*, 454 F.3d 131, 144 (3d Cir. 2006); *Ferensic v. Birkett*, 501 F.3d 469, 477–80 (6th Cir. 2007).

287. *United States v. Holloway*, 971 F.2d 675, 679 (11th Cir. 1992).

288. *United States v. Owens*, 682 F.3d 1358, 1359 (11th Cir. 2012) (memorandum) (Barkett, J., dissenting from denial rehearing en banc) (“Having been given the opportunity to change our court’s position that appellate courts are never permitted to review for abuse of discretion the exclusion of expert testimony regarding the reliability of eyewitness identifications, we should avail ourselves of it. That isolated position, established thirty years ago, conflicts with all of the other circuits and all but five of the states that have considered the question.”).

the centrality of the eyewitness identification evidence, the failure to call an expert was a prejudicial “fail[ure] to render reasonable professional assistance.” The Sixth Circuit explained, in a 2007 ruling finding exclusion of the defense eyewitness expert to be unconstitutional, that the significance of an expert’s testimony “cannot be overstated” because, without it, a jury has “no basis beyond defense counsel’s word to suspect the inherent unreliability” of an eyewitness identification.²⁸⁹

The Third Circuit Court of Appeals convened a task force on eyewitness identifications that issued a detailed report in 2019 that surveyed scientific research regarding eyewitness evidence.²⁹⁰ The report was not binding, but rather made recommendations to law enforcement, advocates, and courts, including recommendations regarding conducting lineups, videotaping lineups, documenting witness confidence, blind administration, as well as continuing education materials for lawyers. The task force did not make recommendations concerning amending jury instructions, noting that a separate Third Circuit committee considers model jury instructions for the court. Those instructions already include recommendations for a range of additions to the *Telfaire* instructions.²⁹¹

State Court Rulings

Several state courts have modified the federal due process rule, relying on the research that has developed in the intervening decades since *Manson*. Several state courts have adopted different factors in a “refinement” of the *Manson* and *Biggers* factors.²⁹² Additional state courts, however, have rejected the *Manson* test, substituting a new framework based on scientific research.²⁹³ Still additional states have adopted changes to jury instructions and expert evidence regarding eyewitnesses.

289. *Ferensic*, 501 F.3d at 477–80.

290. *Report of the United States Court of Appeals for the Third Circuit Task Force on Eyewitness Identifications*, *supra* note 27.

291. Comm. on Model Crim. Jury Instructions for the Third Circuit, Model Criminal Jury Instructions § 4.15 (2017), *supra* note 283.

292. *See, e.g.*, *State v. Ramirez*, 817 P.2d 774, 780–81 (Utah 1991) (altering three “reliability” factors to focus on effects of suggestion), *abrogated by* *State v. Lujan*, 459 P.3d 992 (2020); *State v. Marquez*, 967 A.2d 56, 69–71 (Conn. 2009) (adopting detailed criteria for assessing suggestion); *Brodes v. State*, 614 S.E.2d 766, 771 & n.8 (Ga. 2005) (rejecting use of eyewitness certainty); *State v. Hunt*, 69 P.3d 571, 576 (Kan. 2003) (adopting five factor “refinement” of federal due process test); *State v. Almaraz*, 301 P.3d 242, 253 (Idaho 2013) (“[B]y outlining the system and estimator variables that research has convincingly shown to impact the reliability of eye-witness identification, we hope to provide guidance to lower courts applying the test. . .”).

293. *State v. Henderson*, 27 A.3d 872, 877 (N.J. 2011) (finding scientific evidence shows *Manson* test should be revised); *State v. Lawson*, 291 P.3d 673, 678 (Or. 2012) (en banc) (revising *Manson* test in light of scientific evidence); *Commonwealth v. Gomes*, 22 N.E.3d 897, 909 (Mass. 2015) (finding scholarly research should inform identification instructions); *Young v. State*, 374 P.3d 395,

The most notable rulings have been by the Supreme Courts of Alaska (*Young v. Alaska*, 2016), Connecticut (*State v. Guilbert*, 2012), Hawaii (*State v. Cabagbag*, 2012), New Jersey (*State v. Henderson*, 2011), Oregon (*State v. Lawson*, 2012), Utah (*State v. Clopten*, 2009), and Wisconsin (*State v. Dubose*, 2005).²⁹⁴

The first prominent judicial ruling involving a wholesale reconsideration of the federal due process framework was the New Jersey Supreme Court ruling in *State v. Henderson*. The court-appointed special master “evaluate[d] scientific and other evidence about eyewitness identifications[,] . . . presided over a hearing that probed testimony by seven experts and produced more than 2,000 pages of transcripts along with hundreds of scientific studies,” and then issued a detailed report.²⁹⁵ The decision set out a framework, grounded in scientific research, in which pretrial hearings must examine eyewitness identification evidence to assess its reliability.²⁹⁶ A defendant must show evidence of unreliability, and the state must counter with evidence of reliability. In response, the judge may consider remedies, including jury instructions at trial.

In 2016, the Alaska Supreme Court revised the *Manson* test and adopted a more detailed framework, asking judges to focus on a range of variables that affect the reliability of eyewitness evidence, as in the *Henderson* decision, and adjudicating these questions pretrial, including with the benefit of expert testimony to explain how those variables may apply to different factual situations.²⁹⁷ If the defendant meets this burden, “the trial court should suppress the evidence—both the pretrial identification and any subsequent in-court identification by the witness.” If the defendant does not, however, then “the court should admit the evidence and provide the jury with an instruction appropriate to the context of the case.” The approach, thus, resembles the type of functional framework adopted in New Jersey. In 2018, in *State v. Harris*, the Connecticut Supreme Court adopted a framework much like that in New Jersey (noting the overlap with that approach), holding expert testimony may be valuable, and “it may be appropriate for the trial court to craft jury instructions to assist the jury in its consideration of this issue.”²⁹⁸

427 (Alaska 2016) (urging trial courts to incorporate evolving scientific understanding to supplement test variables).

294. See, e.g., *Young*, 374 P.3d at 427; *State v. Guilbert*, 49 A.3d 705 (Conn. 2012); *State v. Cabagbag*, 277 P.3d 1027, 1035–38 (Haw. 2012) (holding that when identification is central issue, court must, at defendant’s request, give jury instruction about factors affecting reliability); *Henderson*, 27 A.3d 872; *State v. Lawson*, 291 P.3d 673 (Or. 2012); *State v. Clopten*, 223 P.3d 1103 (Utah 2009); *State v. Dubose*, 699 N.W.2d 582 (Wis. 2005). The *Cabagbag* ruling prompted model jury instructions in Hawaii. *State v. Kaneaiakala*, 450 P.3d 761, 774 (Haw. 2019) (holding that when applicable, jury must receive instructions to consider impact of suggestive procedures on identification reliability).

295. *Kaneaiakala*, 450 P.3d at 877.

296. *Henderson*, 27 A.3d. at 894–96, 925–26 (noting research shows “that memory is a constructive, dynamic, and selective process”).

297. *Young*, 374 P.3d at 413.

298. *Harris*, 191 A.3d at 134.

Adopting a slightly different approach, the Hawaii Supreme Court ruled in 2019 that “courts must, at minimum, consider any relevant factors set out in the Hawaii Standard Instructions governing eyewitness and show-up identifications, as may be amended.” Thus, the court did not set out a series of factors, as the New Jersey court did, but rather set out jury instructions that may be changed over time.²⁹⁹ Tracking the New Jersey focus on system variables, the court also emphasized the importance of suggestion: “[T]rial courts must also consider the effect of the suggestiveness on the reliability of the identification in determining whether it should be admitted into evidence.”

In 2015, the Massachusetts Supreme Judicial Court “review[ed] the scholarly research, analyses by other courts, amici submissions,” and the report by a Massachusetts Supreme Judicial Court Study Group on Eyewitness Identification.³⁰⁰ The court recommended judges provide a set of more concise jury instructions on eyewitness identification evidence. The approach similarly includes a range of research-informed factors, but adopts jury instructions that are more manageable in length.

In contrast to these approaches, which rely on a multifactored standard for admissibility and instructions for educating jurors, the Oregon Supreme Court has endorsed review of reliability of eyewitness evidence relying on a general Rule 403 analysis under the Oregon Rules of Evidence. The Oregon court identified two problems with the prior framework: The “threshold requirement of suggestiveness inhibits courts from considering evidentiary concerns” and the “inquiry fails to account for the influence of suggestion on evidence of reliability.”³⁰¹

Other state courts have made more selective changes, such as rejecting eyewitness confidence as a factor to be considered when assessing the validity of an eyewitness identification,³⁰² or requiring additional jury instructions.³⁰³

299. *Kaneaiakala*, 450 P.3d at 777.

300. *Commonwealth v. Gomes*, 22 N.E.3d 897, 900 (Mass. 2015).

301. *State v. Lawson*, 291 P.3d 673, 748 (Or. 2012).

302. *State v. Harris*, 191 A.3d 119, 134 (Conn. 2018). 191 A.3d at 134 (expanding protections for suggestiveness of eyewitness identification procedure); *State v. Almaraz*, 301 P.3d 242, 253 (Idaho 2013) (naming estimator and system variables to be considered in applying *Manson* test); *People v. Adams*, 423 N.E.2d 379, 383–84 (N.Y. 1981) (concluding where witness makes identification under influence of suggestive procedure and there is no independent source that suggests defendant is perpetrator, court should exclude evidence); *State v. Ramirez*, 817 P.2d 774, 781 (Utah 1991) (modifying *Manson* test to remove level of certainty as factor), *abrogated by State v. Lujan*, 459 P.3d 992 (2020); *State v. Hunt*, 69 P.3d 571, 572 (Kan. 2003) (refining *Neil v. Biggers*, 409 U.S. 188 (1972), approach by adding factors from *Ramirez*); *State v. Discola*, 184 A.3d 1177 (Vt. 2018) (rejecting witness certainty as factor in assessing reliability).

303. *State v. Kaneaiakala*, 450 P.3d 761, 777–78 (Haw. 2019) (holding when identification has been procured through suggestive procedure or when central to case, jury must be instructed to consider potential impact of suggestive procedures on reliability of identification).

In addition, several courts have limited the use of courtroom identifications.³⁰⁴ The Massachusetts Supreme Judicial Court and the Connecticut Supreme Court have ruled that no in-court identification is permitted if an out-of-court identification was suppressed as unduly suggestive.³⁰⁵ The Massachusetts Supreme Judicial Court has also ruled that first-time courtroom identifications should not normally be permitted.³⁰⁶ Other courts adopt a burden-shifting approach toward in-court identifications.³⁰⁷

State Eyewitness Evidence Statutes

Lawmakers in almost half of the states have required law enforcement adoption of best practices for eyewitness evidence. State legislation has required the study of eyewitness procedures, development of model policies, or increasingly, outright adoption of eyewitness identification practices by law enforcement. To date, twenty-five states have enacted legislation requiring law enforcement adoption of written eyewitness identification procedures.³⁰⁸ These statutes were often enacted to ensure uniformity in adoption of best practices. Most of those

304. For a proposal that prior identifications should be admitted, but in-court identifications and statements of confidence should not, see Brandon L. Garrett, *Judging Eyewitness Evidence*, 104 *Judicature* 30, 34–35 (2020), <https://perma.cc/5ETW-QUV5>.

305. See *Commonwealth v. Johnson*, 45 N.E.3d 83, 92–94 (Mass. 2016) (reasoning that “a subsequent in-court identification cannot be more reliable than the earlier out-of-court identification, given the inherent suggestiveness of in-court identifications and the passage of time”); *State v. Dickson*, 141 A.3d 810, 817 n.2 & n.3 (Conn. 2016) (“[I]n cases in which identity is an issue, in-court identifications that are not preceded by a successful identification in a nonsuggestive identification procedure implicate due process principles and, therefore, must be prescreened by the trial court.” (footnotes omitted)).

306. See *Commonwealth v. Crayton*, 21 N.E.3d 157, 169 (Mass. 2014) (“Where an eyewitness has not participated before trial in an identification procedure, we shall treat the in-court identification as an in-court showup, and shall admit it in evidence only where there is ‘good reason’ for its admission.”).

307. See *State v. Hickman*, 330 P.3d 551, 568 (Or. 2014), *modified on reconsideration*, 343 P.3d 634 (2015) (“Courts considering the admissibility of first-time in-court identifications generally have placed the burden of seeking a prophylactic remedy on the defendant.”); *United States v. Domina*, 784 F.2d 1361, 1369 (9th Cir. 1986), *cert. denied*, 479 U.S. 1038 (1987) (noting that district court’s denial of request for in-court lineup will only be overturned if in-court identification procedures were so suggestive and conducive to irreparable misidentification as to amount to denial of due process).

308. See Cal. Penal Code § 859.7 (West 2022); Colo. Rev. Stat. § 16-1-109 (2022); Conn. Gen. Stat. § 54-1p (2021); Fla. Stat. § 92.70 (2021); Ga. Code Ann. § 17-20-2 (2021); Hawaii Rev. Stat. 801k (2022); 725 Ill. Comp. Stat. 5/107A-2 (2021); Kan. Stat. Ann. § 22-4619 (2021); La. Code Crim. Proc. Ann. Arts. 251–53 (2021); Md. Code Ann., Pub. Safety § 3-506 (West 2021); Minn. Stat. § 626.8433 (2021); Neb. Rev. Stat. § 81-1455 (2021); Nev. Rev. Stat. § 171.1237 (2021); N.H. Rev. Stat. Ann. § 595-C:2 (2022); N.M. Stat. Ann. § 29-3B-3 (2022); N.Y. Crim. Proc. Law § 60.25 (McKinney 2021); N.C. Gen. Stat. § 15A-284.52 (2021); Ohio Rev. Code Ann. § 2933.83 (West 2021); Okla. Stat. Tit. 22, § 21 (2021); Tex. Code Crim. Proc. Ann. Art. 38.20 (West 2021); Utah Code Ann. § 77-8-4 (West 2021); Vt. Stat. Ann. Tit. 13, § 5581 (2022); Va. Code Ann. § 19.2-390.02 (2021); W. Va. Code §§ 62-1E-1 to -2 (2021); Wis. Stat. § 175.50 (2022).

Reference Guide on Eyewitness Identification

states regulate the eyewitness identification procedures to be used, including blinded lineups (see above section titled “Communication during lineups: The importance of blinded lineup administration”), clear written instructions, and documenting the confidence of an eyewitness.³⁰⁹ In addition, seventeen states have revised their jury instructions, departing from the *Telfaire* model or models reciting factors from *Manson v. Brathwaite*.³¹⁰

For example, Massachusetts developed model jury instructions regarding eyewitness identifications that include both preliminary language, to be read prior to opening statements, and final instructions.³¹¹ The preliminary language advises the jury that “[t]he mind does not work like a video recorder” and cautions that factors during and after the observed event can alter an individual’s memory of that event. The final instructions are far more detailed, describing identification as a three-stage process: perception of an event, storage of information about the event in one’s mind, and later recollection of that stored information.³¹² The instructions caution that a variety of factors, each of which is listed in the instructions themselves, may affect accuracy of a memory at any stage of the process. In selecting the factors to be included in its revised jury instructions, Massachusetts relied on “a near consensus [standard] in the relevant scientific community.”³¹³ The analysis suggested five such principles (although without referring to factors that affect the validity of visual perceptual experience):

1. human memory does not operate like a video recording that a person can replay to recall what happened;
2. a witness’s level of confidence in an identification may not indicate its accuracy;
3. high levels of stress can reduce the likelihood of making an accurate identification;
4. information from other witnesses or outside sources can affect the reliability of an identification and inflate an eyewitness’s confidence in the identification; and

309. See Albright & Garrett, *supra* note 253, at 511, Appendix A.

310. Those states are Alaska, Connecticut, Florida, Georgia, Hawaii, Kansas, Maine, Maryland, Massachusetts, Missouri, New Jersey, New York, North Carolina, Ohio, Pennsylvania, Utah, and Virginia.

311. Mass. Sup. Jud. Ct., Model Jury Instructions on Eyewitness Identification: Preliminary/Contemporaneous Instruction 2 (2015); Mass. Sup. Jud. Ct., Model Jury Instructions on Eyewitness Identification: Model Eyewitness Identification Instruction 2 (2015) (establishing jury instruction to be given where jury heard evidence affirmatively identifying defendant and such identification is contested).

312. Model Jury Instructions on Eyewitness Identification: Model Eyewitness Identification Instruction 2 (Mass. Sup. Jud. Ct. 2015).

313. *Id.* at 2–10.

5. viewing the same person in multiple identification procedures may increase the risk of misidentification.³¹⁴

We note that there is not strong evidence that these jury instructions fully accomplish their aim to better educate jurors regarding the potential strengths and weaknesses of eyewitness evidence, including because jurors place such powerful weight on courtroom identifications and courtroom expressions of confidence.³¹⁵ Indeed, research has not found that the instructions adopted in New Jersey, or the more concise Massachusetts instructions, are effective in helping jurors distinguish between high- and low-confidence initial identifications.³¹⁶ More recent research suggests that instructions explaining *why* it is important to place weight on initial confidence, and not courtroom identifications and courtroom confidence, may help jurors make that important distinction.³¹⁷

Police Practice Recommendations

In recent years, professional policing organizations, such as the IACP, Major Cities Chiefs Association, and the CALEA, have taken active roles in promoting best practices, including through model policies.³¹⁸ The Third Circuit task force emphasized that results of research on eyewitness procedures must be translated to law enforcement personnel with proper guidelines for maximum effectiveness. The U.S. Department of Justice adopted a policy in 2017, applying to all federal law enforcement agencies, that requires blind administration of lineups, among other procedures.³¹⁹ Twenty-six states and the federal government have voluntarily adopted such policies regulating eyewitness identifications.³²⁰ The model policies generally require blinded lineups (see above section titled “Communication during lineups: The importance of blinded lineup administration”),

314. Ralph D. Gants & Erik N. Doughty, *Where Science Conflicts with Common Sense: Eyewitness Identification Reform in Massachusetts*, 79 Alb. L. Rev. 1617, 1625–26 (2017) (quoting *Commonwealth v. Gomes*, 22 N.E.3d 897, 903, 909 (Mass. 2015)).

315. See generally, Garrett et al., *supra* note 33.

316. *Id.*

317. Garrett et al., *supra* note 34.

318. See U.S. Dep’t of Just./Int’l Ass’n of Chiefs of Police, National Summit on Wrongful Convictions: Building A Systemic Approach to Prevent Wrongful Convictions 13–14 (2013), <https://perma.cc/ML3T-2ZJE>; Comm’n on Accreditation for L. Enf’t, CALEA L. Enf’t Agency Standards § 42.2.11 (2010); Int’l Ass’n of Chiefs of Police Pol’y Ctr., Model Pol’y: Eyewitness Identification 1–4 (2016).

319. Procedures for Conducting Photo Arrays, *supra* note 7.

320. The states are Arkansas, Colorado, Connecticut, Delaware, Florida, Georgia, Kansas, Kentucky, Louisiana, Maine, Maryland, Michigan, Minnesota, Montana, Nebraska, Nevada, New Hampshire, New Jersey, New Mexico, New York, Rhode Island, Utah, Vermont, Virginia, Washington, and Wisconsin.

clear written instructions, and documenting confidence; many recommend videotaping. Policing organizations have incorporated research recommendations into lineup policies and training; federal support has been important to such efforts.³²¹

National organizations have also set out recommendations for law enforcement regarding incorporation of research into eyewitness identification procedures. The American Law Institute, as part of its Principles of Policing project, adopted principles for law enforcement concerning eyewitness identifications.³²² In 2020, the AP-LS released a scientific review paper summarizing research in the field and making new recommendations.³²³

Conclusion

The evolving law of eyewitness evidence shows how science can play a pivotal role in our legal system. While constitutional precedent can change slowly, its practical application can be informed by scientific research. Evidentiary practices, like appointment of experts, pretrial hearings and rulings, and jury instructions, can also be informed by scientific research. We have experienced a paradigm shift: Almost every jurisdiction—through judicial rulings, legislation, model policy, and police policy—has embraced scientific research regarding eyewitness evidence. These approaches aim to more accurately collect evidence from eyewitnesses, and to prevent contamination of that evidence through suggestion and other forms of bias.

321. These efforts began in the late 1990s with the National Institute of Justice's Technical Working Group for Eyewitness Evidence. See U.S. Dep't of Just., Nat'l Inst. of Just., *Eyewitness Evidence: A Guide For Law Enforcement*, *supra* note 27 (putting forth recommendations to improve procedures for collection and preservation of eyewitness evidence within criminal justice system).

322. Principles of the Law: Policing ch. 10 (A.L.I. 2022).

323. Wells et al., *Policy and Procedure Recommendations* (2020), *supra* note 27, at 3.

Glossary of Terms

accuracy. The probability that the identification is of the perpetrator. Estimated by predictive variables in real casework and commonly measured in known source tests as the number of correct identifications relative to the total number of identifications (correct + incorrect).

attentional hijacking. The condition in which a highly salient stimulus directs the focus of attention away from an intended target of attention.

bias. To evaluate something in a prejudicial manner. The meaning here is not pejorative: It refers to prior knowledge or experience that is used to infer the meaning of an incomplete or ambiguous sensory event.

binary classification. A cognitive process in which the individual must decide between two discrete alternatives, as in the lineup participant is a match or a nonmatch to memory of the perpetrator of a crime.

blinded test. (Also known as “single-blinded test.”) A classification task in which the witness is prevented (hence “blinded”) from knowing the status of lineup participants. See double-blinded test.

categorical decision. A decision in favor of one of a set of discrete alternatives, such as classification of a lineup participant as a match or a non-match.

categorical perception. The experience in which different sensory stimuli of the same category are perceived similarly, which entails loss of sensitivity to individual sensory differences within a category.

center of gaze. The region of visual space corresponding to the portion of the retina with the highest density of photoreceptors; the region with the highest visual acuity.

color blindness. A genetically determined condition in which an individual has fewer than three types of color-specific photoreceptors, which limits the range of discriminable colors.

confidence. A cognitive state and expression that reflects perceived certainty of a decision, as in perceived certainty that one’s identification of the perpetrator is correct.

confidence inflation. The increase in perceived certainty in a decision that follows from reinforcement and perceived corroboration by others.

contextual variables. Variables associated with a signal event, as in environmental, perceptual, or cognitive states that are correlated with the experience of witnessing a crime and identifying a perpetrator. These variables may be useful in assessing the accuracy of the identification.

contrast sensitivity. A quantified measure of an observer’s sensitivity to a difference between two sensory stimuli, such as different intensities of light, which bears on the ability to detect and discriminate patterns.

cross-race effect. A well-established perceptual phenomenon in which the ability of an observer to distinguish between faces of members of their own race exceeds the ability to distinguish between faces of members of a different race. The effect is believed to be a consequence of perceptual learning resulting from predominant exposure of individuals to same-race faces during critical stages of development. It has the obvious potential to yield differential error rates in same-race versus cross-race eyewitness identifications.

decision criterion. The threshold value of a decision variable that an observer uses to distinguish two different classifications of a sensory stimulus, as in the specific degree of perceptual similarity between a lineup face and a face stored in memory that distinguishes between a match versus a non-match decision.

decision variable. The measured or computed variable that an observer uses to make a classification decision, as in the similarity between the appearance of a lineup face and the eyewitness's memory of the face of the culprit.

declarative memory. Long-term memory for facts and events that can be retrieved at will and brought into conscious awareness.

degree of visual angle. The angular extent of visual space from the nodal point of the eye, which is commonly used to refer to the spatial extent of an object as optically projected onto the back of the eye. In the context of object recognition, it is used to determine the ability to resolve features in the projected image.

diagnosticity ratio (DR). The ratio of the probability of correct identification of a known culprit to the probability of incorrect identification of a known innocent suspect. Used in laboratory studies of eyewitness performance as a measure of the ability of a witness to distinguish the culprit from an innocent person. Not a true measure of discriminability, since the diagnosticity ratio is also influenced by the decision criterion employed by the eyewitness.

discriminability. Reflects the degree to which an instrument of measurement can distinguish the values of two inputs, assessed independently of the decision criterion employed to make the classification decision. In the eyewitness context, this refers to the extent to which the witness can tell the difference between the culprit and an innocent suspect.

double-blinded test. A classification task in which the witness and the administrator of the task are both "blinded" from knowing the status of lineup participants; this is modeled after an element of the scientific method designed to prevent decision bias based on prior knowledge of experimental conditions.

ecological validity. The extent to which variables, results, and conclusions from an experimental study generalize to the real world.

episodic memory. The form of declarative memory that holds information about the sequence or narrative content of memory. Eyewitness memories are of this type.

estimator variable. Contextual factors associated with the viewing conditions and perceptual/cognitive state of the witness at the time of the crime. These include environmental factors such as intensity of illumination and duration of events, as well as observer variables, such as acuity, emotion, and pain. These variables are not controllable in an eyewitness context. However, because they can predictably impair the fidelity of visual perception and memory encoding, they thus have great potential to inform the trier of fact as to the probability that an identification is accurate.

face blindness. Laboratory studies have shown that face recognition ability varies continuously across the human population. At one extreme of this distribution are “face-blind” people who are extremely poor at recognition memory for faces. The degree of face blindness can be quantified by now-standard face recognition tests. Performance on such tests is an estimator variable that can be used for quantitative prediction of the accuracy of eyewitness reports. Contrasted with face super-recognizers.

face super-recognizer. Laboratory studies have shown that face recognition ability varies continuously across the human population. At the extreme opposite end of the continuum from face-blind people are face super-recognizers, who excel at recognition memory for faces. The degree of performance can be quantified by now-standard face recognition tests, which can serve as quantitative predictors of the accuracy of eyewitness reports.

fair lineup. A lineup composed of fillers who all bear a similar perceptual relationship to the witness’s reported description of the culprit. Fair lineups are less likely to elicit biased identifications and wrongful convictions of people who stand out from the crowd based on similarity to the witness’s description.

gaze direction. The point in visual space that is the momentary target of the center of gaze, which possesses the highest visual acuity.

illusory conjunctions. A perceptual phenomenon in which briefly viewed attributes of a scene become disjoined from the objects they originate from and conjoined with others, as in a red letter D and a blue number 2 that are perceived as a blue D and a red 2.

inattentional blindness. Failure to perceive an otherwise prominent visual stimulus that results from lack of attentional focus directed to the stimulus.

lineup type. The specific type of visual presentation and behavioral task used to test an eyewitness’s ability to identify the culprit.

long-term memory. Memories that have been consolidated into long-term storage, where they are less subject to forgetting but still modifiable by the acquisition of new information.

luminance contrast. A measure of the contrasting amounts of light reflected (i.e., luminance) from two spatially adjacent features of a visual scene, which is a primary determinant of an observer's ability to perceptually distinguish the features.

memory consolidation. The transfer of perceptual or cognitive information from labile short-term storage into more stable long-term storage, which is a time-dependent process mediated by changes in neuronal connectivity.

memory encoding. The transfer of perceptual or cognitive experiences into a form storable by the brain, which underlies the ability to learn from those experiences.

memory plasticity. Memory is modifiable over time, to a high degree during acquisition and to a lesser degree once consolidated. Memory systems are more plastic during critical stages of human development, but retain some plasticity throughout life.

memory retention interval. The amount of time that passes between initial acquisition of a new memory and a subsequent test of the fidelity of that memory. Within that retention window, memories predictably fade because of forgetting and can be modified by intrusion of new information.

memory retrieval. The process of bringing a stored piece of information back to a state where it can be used to guide decisions and behavior. Retrieval is commonly elicited by an external sensory stimulus or by an internal association.

memory storage. The state in which memories are maintained for the long term and accessible for retrieval at a later time. Memories are less subject to forgetting during this period but can still be modified by acquisition of new information.

misdirection. Direction of the focus of visual attention away from the location of a behaviorally relevant object or feature. A prominent tool used in magical tricks that demonstrates the easy fallibility of visual perceptual experiences and the inappropriate assertion of confidence in such experiences. This phenomenon underlies the "weapon focus" effect that impairs accuracy of eyewitness identification.

myopia. A visual pathology in which the optics of the eye are misaligned, such that the projected pattern of light on the back surface of the eye is not well focused. Although correctable with external lenses or corneal surgery, this pathology is common enough in the human population, to varying degrees, that it could easily limit the accuracy of eyewitness identification.

noise. Random or irrelevant elements that interfere with detection of coherent and informative signals. In human sensory systems, this interference is frequent and can take the form of distortion of signals at the sensory periphery, such as optical impairment, or by diffraction of light by fluid in the eye. Other sources

of noise include occluding elements of a visual scene or attentional distractions. The result is reduced fidelity of sensation and perceptual uncertainty.

object surface reflectance. The amplitude and chromatic spectrum of white light reflected from the surface of an object. This reflectance is a physical property of the surface itself, which interacts with the amplitude and chromatic spectrum of light illuminating the scene.

operating characteristic. A set of parameters that define the performance of an instrument based on empirical assessment and knowledge of its underlying mechanism. Generally, these describe performance in terms of the validity and reliability of the output for a given type of input. Quantitative performance descriptions of this sort offer a useful systems approach to understanding the abilities and weaknesses of eyewitness reports.

perception. The process by which evanescent sensations are linked to environmental cause and made enduring and coherent through the assignment of meaning, utility, and value.

perceptual improvisation. Unconscious search through memory for information (probabilities of the accuracy of previously acquired knowledge, beliefs, and hypotheses) that enables rapid identification, recognition, and spatial understanding of novel and otherwise unrecognizable stimuli.

perceptual learning. Long-term experience-dependent changes in the way that people perceive sensory stimuli, which can enhance or distort perceptual experience. Apropos of eyewitness identification, the cross-race effect, which comes about through extended selective exposure to faces of people of one's own race, is a consequence of perceptual learning.

priors. Shorthand for “prior probability,” which refers to the probability of accuracy for knowledge, beliefs, or hypotheses about the world that an observer holds prior to a signal event, such as witnessing a crime. Used synonymously with prejudices or biases, priors enable an observer to unconsciously fill in blanks in the presence of perceptual uncertainty.

receiver operating characteristic (ROC). A standard modeling approach to the evaluation of performance on a classification task, drawn from signal detection theory. In the eyewitness context, the ROC is generally presented as a curve that plots the probability of correct detection of a target (the culprit) against the probability of incorrect detection of a non-target (an innocent suspect). The area under the ROC curve is a single-value measure of the ability to discriminate the culprit from an innocent suspect, which has the advantage of being independent of the decision criterion used by the eyewitness. Use of this approach has transformed analysis of data from laboratory studies of factors that affect eyewitness performance, in part because it yields a true measure of discriminability.

recognition memory. The ability to determine whether a present sensory stimulus (a “cue”) is the same as a stimulus previously seen and memorized, and subsequently retrieved from memory. In the eyewitness context, a positive outcome from this similarity assessment is the event that leads to an assertion of identification.

reliability. A property of an instrument of measurement (e.g., an eyewitness) that reflects the degree to which it yields the same result every time it is used to measure the same thing. Often used inappropriately to refer to the validity of an instrument of measurement.

retina. The sheet of neuronal tissue that lines the back of the eye. Light reflected from surfaces in the visual environment passes through the lens of the eye and is projected onto the retina, which includes the photoreceptors that are responsible for phototransduction (conversion of luminous energy into neuronal energy). Numerous optical and neuronal noise sources can create visual uncertainty at this stage of processing, undermining the accuracy of perceptual experience.

semantic memory. The form of declarative memory that holds information about meanings of experiences and acquired knowledge of the world.

sensation. The consequence of spatially and temporally patterned stimulation of biological sensory receptors. Antecedent to and contrasted with perception, through which sensations are linked to environmental cause.

sequential lineup. A lineup identification procedure in which each face in the lineup is presented to the witness in isolation, and the complete set is presented in a temporal sequence. Believed to promote absolute lineup comparisons.

short-term memory (working memory). The memory system that has limited capacity for volume of content and duration of storage, but where content is actively maintained and readily available for use to guide decisions and actions (i.e., it is “working” memory). Short-term memory is highly labile storage, subject to distortion or loss in the presence of distractions. Extended maintenance of content requires consolidation into long-term memory storage.

signal detection. From signal detection theory, which is a theoretical framework for understanding the processes and consequences of decision-making in the presence of uncertainty. Used for decades as an efficient means to analyze behavioral responses to sensory stimulation.

simultaneous lineup. A lineup identification procedure in which all faces in the lineup are presented to the witness at the same time.

source memory failure. Loss of the ability to remember how one acquired a specific memory, while the memory itself is intact. A common consequence of this failure is that the source of the memory is wrongly attributed based on priors and may incorrectly modify other memories. This is a memory

system analogue of the introduction of biases to resolve uncertainty in visual perception.

system variables. Procedures and states that may influence the accuracy of an eyewitness report and are potentially under the control of the criminal justice system, since they bear on activities that occur after the witnessed events. Includes lineup type and blinded versus nonblinded lineup administration.

target absent lineup. A type of lineup procedure that does not include the culprit (“target”) as one of the lineup participants. Commonly refers to an empirical practice in laboratory studies of eyewitness performance, which is used to assess the probability that witnesses mistakenly identify an innocent suspect. Results (probabilities of incorrect identifications) are compared to those obtained using a target present lineup. The ratio of correct to incorrect identification probabilities is the diagnosticity ratio.

target present lineup. A type of lineup procedure that includes the culprit (“target”) as one of the lineup participants. Commonly refers to an empirical practice in laboratory studies of eyewitness performance, which is used to assess the probability that witnesses correctly identify the culprit. Results (probabilities of correct identifications) are compared to those obtained using a target absent lineup. The ratio of correct to incorrect identification probabilities is the diagnosticity ratio.

task irrelevant information. Contextual information associated with a visual or cognitive task that is not required to solve the task but is at risk of biasing the outcome. Apropos of eyewitness identification, task irrelevant information can refer to reports from journalists, family, friends, and legal counsel that may artifactually influence a witness’s degree of certainty in their identification. Contrasted with task relevant information, which is needed to solve the task.

uncertainty. In the eyewitness context, the state of being uncertain about a measured piece of sensory information, as a result of noise and ambiguity. The common consequence is that uncertainty is resolved by completion of perceptual experience based on prior knowledge or beliefs.

unfair (biased) lineup. A lineup in which the chosen fillers possess different degrees of similarity of appearance to the witness’s description of the culprit. Unfair lineups can bias the choices made by a witness because some faces stand out from the crowd, which may lead to wrongful investigation, prosecution, and conviction.

validity. The property of an instrument of measurement that reflects the degree to which it yields the result that it was intended to measure. In the eyewitness context, this generally refers to the measured probability that eyewitnesses will correctly identify the culprit, which is typically obtained from laboratory studies in which ground truth is known.

Reference Guide on Eyewitness Identification

visual acuity. A measure of the clarity and sharpness of vision, empirically assessed by the threshold ability to resolve points in an image at varying degrees of angular separation. A measure of an eyewitness's visual acuity is essential to predicting the accuracy of an identification.

visual crowding. Reduction of the probability of correctly detecting and perceiving a target stimulus when it is viewed in a dense arrangement of other stimuli. In the eyewitness context, this is manifested as increasing limits on the ability to correctly perceive the actors and events of a crime as a function of the density of actors and objects in the scene.

visual performance. A generic term for the quantifiable ability of an observer to accurately detect and discriminate visual stimuli.

visual search. A well-studied perceptual phenomenon in which the amount of time it takes to detect a specific object amid a set of related objects is dependent on the presence or absence of distinctive features. In the presence of distinctive features, such as a unique color, immediate perceptual “pop-out” occurs. If, on the other hand, the object is defined by a unique conjunction of features (for example, a single red letter T in a field of green Ts and both red and green Ls), time to target object detection is predictably delayed because the visual system must “search” serially through the scene until it is encountered. This attentional delay has the potential to markedly increase uncertainty in time-limited eyewitness events.

Reference Guide on Statistics and Research Methods

DAVID H. KAYE AND HAL S. STERN

David H. Kaye, M.A., J.D., is Regents Professor Emeritus, Arizona State University Sandra Day O'Connor College of Law and School of Life Sciences, and Distinguished Professor of Law and Academy Professor Emeritus, Pennsylvania State University School of Law.

Hal S. Stern, Ph.D., is Distinguished Professor, Department of Statistics, University of California, Irvine.

Authors' Note: Research for this reference guide was completed in 2023. Professor David A. Freedman co-authored the first three editions of this reference guide. His writing and thinking are evident throughout this edition as well.

CONTENTS

Introduction, 466

Admissibility and Weight of Statistical Studies, 467

Varieties and Limits of Statistical Expertise, 467

Procedures That Enhance Statistical Testimony, 469

Maintaining Professional Autonomy, 469

Disclosing Limitations and Other Analyses, 470

Disclosing Data and Analytical Methods Before Trial, 470

How Have the Data Been Collected?, 471

Is the Study Designed to Investigate Causation?, 472

Types of Studies, 472

Randomized Controlled Experiments, 474

Observational Studies, 475

Generalizing the Results, 478

Descriptive Surveys and Censuses, 479

What Method Is Used to Select the Units?, 479

A sampling frame, 480

Selection bias, 480

Of the Units Selected, Which Provide Measurements?, 482

Individual Measurements, 483

Is the Measurement Process Reliable?, 483

Is the Measurement Process Valid?, 485

Are the Measurements Recorded Correctly?, 487

What Does It Mean to Be Random?, 488

- How Have the Data Been Presented?, 488
 - Are Rates or Percentages Properly Interpreted?, 489
 - How Big Is the Base of a Percentage?, 489
 - Have Appropriate Benchmarks Been Provided?, 489
 - Have the Data Collection Procedures Changed?, 490
 - Are the Categories Appropriate?, 491
 - What Comparisons Are Made?, 493
 - Is an Appropriate Measure of Association Used?, 494
 - Does a Graph Portray Data Fairly?, 496
 - How Are Trends Displayed?, 496
 - How Are Distributions Displayed?, 497
 - Is an Appropriate Measure Used for the Center of a Distribution?, 499
 - Is an Appropriate Measure of Variability Used?, 500
- What Inferences Can Be Drawn from the Data?, 502
 - Estimation, 504
 - What Estimator Should Be Used?, 504
 - What Is the Standard Error?, 505
 - What Is the Confidence Interval?, 506
 - The normal curve and large samples, 506
 - Other situations, 509
 - How Big Should the Sample Be?, 510
 - What Are the Technical and Interpretive Difficulties with Confidence Intervals?, 511
 - p -values, Significance Levels, and Hypothesis Tests, 514
 - What Is the p -value?, 514
 - Is a Difference Statistically Significant?, 517
 - Recent Emphasis on the Limitations of p -values, 519
 - Tests or Interval Estimates?, 520
 - Is the Sample Statistically Significant?, 521
 - Evaluating Hypothesis Tests, 522
 - What Is the Power of the Test?, 522
 - What About Small Samples?, 523
 - One Tail or Two?, 523
 - How Many Tests Have Been Done?, 524
 - What Are the Rival Hypotheses?, 526
 - Bayesian Statistical Methods and Posterior Probabilities, 527
- Correlation and Regression, 529
 - Scatter Diagrams, 529
 - Correlation Coefficients, 530
 - Is the Association Linear?, 532
 - Do Outliers Influence the Correlation Coefficient?, 532
 - Does a Confounding Variable Influence the Coefficient?, 533
 - Regression Lines, 534
 - What Are the Slope and Intercept?, 534
 - What Is the Unit of Analysis?, 536

Reference Guide on Statistics and Research Methods

- Statistical Models, 538
- Data Science and Statistical Machine Learning, 543
 - What Is Data Science?, 544
 - What Is Machine Learning?, 544
 - What Statistical Questions Arise with Machine Learning Studies?, 547
 - Is the Dataset Appropriate and of Sufficient Quality?, 548
 - Is the Predictor or Classifier Robust?, 548
 - Is the Predictor or Classifier Too Opaque?, 549
- Appendix: Conditional Probability and Bayes' Rule, 550
 - What Do Probabilities Apply To?, 550
 - What Are Conditional Probabilities?, 551
 - What Is Bayes' Rule?, 551
- Glossary of Terms, 555
- References on Statistics and Research Methods, 575
 - Nontechnical Surveys, 575
 - General References, 575

FIGURES

- 1 and 2. Manipulating the scale of a graph, 497
3. Histogram showing how frequently various numbers of heads appeared in 50 batches of 10 tosses of a quarter, 498
4. Confidence coefficients for CIs of ± 1 , 2, and 3 standard errors of a normally distributed estimator, 508
5. Plotting points in a scatter diagram, 530
6. Scatter diagram for income and education: men ages 25 to 34 in Kansas, 530
7. The correlation coefficient measures the sign of a linear association, and its strength, 531
8. A strong nonlinear association with a correlation coefficient close to zero, 532
9. The correlation coefficient can be distorted by outliers, 533
10. The regression line for income on education and its estimates, 534
11. Scatter diagram for income and education, with the regression line indicating the trend, 535
12. Turnout rate for the white candidate plotted against the percentage of registrants who are white. Precinct-level data, 1982 Democratic primary for auditor, Lee County, South Carolina, 538

TABLES

1. Test Results for Cartridge-case Comparisons, 487
2. Data Used by a Defendant to Refute Plaintiff's False Advertising Claims, 491
3. Home Pregnancy Test Results, 491
4. Test Results for Cartridge-case Comparisons, 492
5. Admissions by Sex, 494
6. Admissions by Sex and College, 496

Introduction

Statistical assessments are prominent in many kinds of legal cases, including anti-trust, employment discrimination, toxic torts, and voting rights cases. This reference guide describes the elements of statistical reasoning. We hope the explanations will help judges and lawyers to understand statistical terminology, to see the strengths and weaknesses of statistical arguments, and to apply relevant legal doctrine. The guide is organized as follows:

- This introduction provides an overview of the field, discusses the admissibility of statistical studies, and offers some suggestions about procedures that encourage the best use of statistical evidence.
- The section titled “How Have the Data Been Collected?” addresses data collection. It explains why the design of a study is the most important determinant of its quality. The section compares experiments with observational studies and surveys with censuses, indicating when the various kinds of study are likely to provide useful results.
- The section titled “How Have the Data Been Presented?” discusses the art of summarizing data. This section considers the mean, median, and standard deviation. These are basic descriptive statistics, and most statistical analyses use them as building blocks. This section also discusses patterns in data that are brought out by graphs, percentages, and tables.
- The section titled “What Inferences Can Be Drawn from the Data?” describes the logic of statistical inference, emphasizing foundations and disclosing limitations. This section covers estimation, standard errors and confidence intervals, p -values, and hypothesis tests.
- The section titled “Correlation and Regression” shows how associations can be described by scatter diagrams, correlation coefficients, and regression lines. Regression is often used in attempts to infer causation from association. This section explains the technique, indicating the circumstances under which it and other statistical models are likely to succeed—or fail.
- Advances in computing speed and the availability of large datasets have made it possible to fit models of greater complexity than those described in the section titled “Correlation and Regression.” The section titled “Data Science and Statistical Machine Learning” describes the developing fields of “data science” and “machine learning” and statistical issues that affect the confidence one can have in the output of machine-learning techniques.
- An appendix discusses the scope of the theory of probability, conditional probabilities, Bayes’ rule, and perspectives on statistical inference.
- The glossary defines statistical terms that may be encountered in litigation, including ones that do not appear in the body of the guide.

Admissibility and Weight of Statistical Studies

Statistical studies suitably designed to address a material issue generally will be admissible under the Federal Rules of Evidence. The hearsay rule rarely is a serious barrier to the presentation of statistical studies, because such studies may be offered to explain the basis for an expert's opinion or may be admissible under the learned treatise exception to the hearsay rule.¹ Most statistical methods applied in litigation are described in textbooks or journal articles and are capable of producing useful results when properly applied. As such, these methods generally satisfy important aspects of the “scientific knowledge” requirement in *Daubert v. Merrell Dow Pharmaceuticals, Inc.*² However, a particular study may use a method that is entirely appropriate but so poorly executed that it should be inadmissible under Federal Rules of Evidence 403 and 702.³ Or, the method may be inappropriate for the problem at hand and thus lack the “fit” referred to in *Daubert*.⁴ Or the study might rest on data of the type not reasonably relied on by statisticians or substantive experts and hence not be suitable as a basis for an opinion under Rule 703. Often, however, the battle over statistical evidence concerns weight or sufficiency rather than admissibility.

Varieties and Limits of Statistical Expertise

For convenience, the field of statistics may be divided into three subfields: probability theory, theoretical statistics, and applied statistics. Probability theory is the mathematical study of outcomes that are governed, at least in part, by chance. Theoretical statistics is about understanding the properties of statistical procedures, including error rates; probability theory plays a key role in this endeavor. Applied statistics draws on both these fields to develop techniques for collecting or analyzing particular types of data.

Statistical expertise is not confined to witnesses with degrees in statistics. Because statistical reasoning underlies many kinds of empirical research, scholars in a variety of fields—including biology, business, economics, epidemiology, medicine, political science, psychology, and sociology—are exposed to statistical ideas, with an emphasis on the methods most important to the discipline. The diffusion of statistical concepts and methods across so many fields raises the question

1. See generally 2 McCormick on Evidence §§ 321, 324.3 (Robert P. Mosteller ed., 8th ed. 2020). Studies published by government agencies also may be admissible as public records. *Id.* § 296.

2. 509 U.S. 579, 589–90 (1993).

3. See *Kumho Tire Co. v. Carmichael*, 526 U.S. 137, 152 (1999) (suggesting that the trial court should “make certain that an expert, whether basing testimony upon professional studies or personal experience, employs in the courtroom the same level of intellectual rigor that characterizes the practice of an expert in the relevant field”).

4. *Daubert*, 509 U.S. at 591.

of who is qualified to conduct and testify to statistical assessments—a statistical methodologist, a substantive scientist, or both? Much depends on context.

If the study involves assembling and then analyzing case-specific data, the choice of which data to examine and how best to model a particular process could require both subject-matter and general statistical expertise. The two types of expertise might be combined in a single individual. A labor economist, for example, should be able to supply a definition of the relevant labor market from which an employer draws its employees. This economist also might be sufficiently educated in statistical tests for group differences to compare the race of new hires to the racial composition of the labor market. When the analysis is as straightforward as the comparison of two proportions, a single substantive expert may suffice.⁵ But further analysis of the possible reasons for the racial disparity might require input from an expert with an understanding of more advanced methods. This expert might be a statistician or an econometrician, or a scientist or engineer who works with other data in a completely different field of study.

The critical question is whether the individual has the education and experience to determine which statistical techniques are appropriate for the task at hand and to apply such techniques with an appreciation of their limitations. Experts who specialize in using statistical methods, and whose professional careers demonstrate this orientation, are most likely to use appropriate procedures and correctly interpret the results. To ascertain the extent to which an expert has this orientation and acumen, one can look to formal education, professional accomplishments, and reputation in the pertinent community of experts. Has the individual studied quantitative methods? Taught them? Used them in research? Which ones? If the expert is or was an academic professional, a university teaching portfolio with courses on statistical methods is a good sign. Ideally, the publication and research record will include studies that use or describe the same or similar methods as those applied (or applicable) to the case at bar.⁶ Invitations from mainstream journal editors to review articles submitted for publication, from academic conference organizers to present research involving statistics, and from government regulatory and research agencies to serve on expert panels that evaluate statistical work are further indications of recognized expertise within a statistical discipline. General membership in most scientific and professional societies requires no special accomplishments, but certain designations, prizes, or awards from such organizations for scholarship are a sign that an

5. Of course, the case could be presented in the form of interlocking testimony from two experts—the labor economist followed by any expert with the appropriate expertise in the method for making the statistical comparison. The substantive knowledge may be of lesser value in selecting a statistical model. Various models might be consistent with the substantive knowledge, and there may be purely statistical criteria for choosing among them.

6. Academic experts are likely to have a long list of publications, but length alone is not the best indication of pertinent expertise. Not all publishers are equal, and every field has a relatively small number of top-tier journals.

expert's work has had an impact. Of course, these factors are merely indicia of degrees of expertise and specialization. Highly competent work can be done by individuals with shorter curricula vitae.

Again, not every case involving computations of probability or statistics necessitates a highly skilled statistical specialist. Medical practitioners and forensic scientists who are not statistical specialists often make statistical assessments of evidence or rely on and present the results of statistical studies. In these situations, it is important to ensure that the witnesses do not exceed the bounds of their statistical expertise. The scientist or practitioner might lack basic information about the studies underlying their testimony. *State v. Garrison*⁷ illustrates the problem. In this murder prosecution involving bitemark evidence, a dentist was allowed to testify that “the probability factor of two sets of teeth being identical in a case similar to this is, approximately, eight in one million,” even though “he was unaware of the formula utilized to arrive at that figure other than that it was ‘computerized.’”⁸ Likewise, laboratory analysts or examiners may have only a limited understanding of the foundations of automated systems for comparing patterns within DNA samples, toolmarks, fingerprints, voice recordings, and the like. Once the formulas or methods have been shown to work well, and when their limitations and proper use are known, the practitioners may be qualified to operate the statistical machinery, as it were, without an in-depth understanding of the routinized or automated statistical or probabilistic method. But the operational training may not qualify them to explain the developmental research. They may have to limit their statistical testimony to an explanation of how they arrived at their estimates and to statements of the extent to which software or hardware is used in practice and whether there are publications on its validity.

Procedures That Enhance Statistical Testimony

Maintaining Professional Autonomy

Ideally, experts who conduct research in the context of litigation should proceed with the same objectivity that would be required in other professional contexts. Thus, experts who testify (or who supply results used in testimony) should conduct the analysis required to address, in a professionally responsible fashion, the issues posed by the litigation.⁹ Questions about the freedom of inquiry accorded

7. 585 P.2d 563 (Ariz. 1978).

8. *Id.* at 566, 568. For other examples, see David H. Kaye et al., *The New Wigmore: A Treatise on Evidence: Expert Evidence* § 12.2 (2d ed. 2011).

9. See Nat'l Research Council Panel, *The Evolving Role of Statistical Assessments as Evidence in the Courts* 164 (Stephen E. Fienberg ed., 1989) [hereinafter NRC Panel] (recommending that the expert be free to consult with colleagues who have not been retained by any party to the litigation and that the expert receive a letter of engagement providing for these and other safeguards).

to testifying experts, as well as the scope and depth of their investigations, may reveal some of the limitations to the testimony.

Disclosing Limitations and Other Analyses

Statisticians analyze data using a variety of methods. To permit a fair evaluation of the analysis that is eventually settled on, the testifying expert can be asked how that approach was developed, whether alternative approaches were considered, and, if so, what the results were.¹⁰ Ethical guidelines for statisticians require them to disclose known and suspected limitations in the data and its analysis.¹¹

Disclosing Data and Analytical Methods Before Trial

The collection of data often is expensive and subject to errors and omissions. Moreover, careful exploration of the data can be time-consuming. To minimize debates at trial over the accuracy of data and the choice of analytical techniques, pretrial-discovery procedures should be used, particularly with respect to the quality of the data and the method of analysis.¹² In some cases, these could include requirements for specifying in detail the study design and analysis that are planned *before* data collection begins.¹³

10. *Id.* at 167; cf. Edith Beerdsen, *Litigation Science After the Knowledge Crisis*, 106 Cornell L. Rev. 529 (2021) (discussing responses to the “replication crisis” in the academic biomedical and behavioral science publication process and suggesting pretrial procedures to make the exercise of “analytical flexibility” in “litigation science” more transparent); Yoav Benjamini, *Selective Inference: The Silent Killer of Replicability*, 2 Harv. Data Sci. Rev. (2020), <https://doi.org/10.1162/99608f92.fc62b261> (noting the problem of presenting or emphasizing selected findings within scientific publications).

11. ASA Comm. on Professional Ethics, *Ethical Guidelines for Statistical Practice*, Feb. 2022, at 3 & 4, <https://perma.cc/P4P6-JU6C>. Invoking an attorney’s professional duty not to intentionally mislead the courts on the facts or the law, the Panel on Statistical Assessments as Evidence insisted that “it is not appropriate for the attorney to seek to avoid such revelation [of alternative forms of data and analyses] by consulting a series of experts without revealing to the experts ultimately retained any prior history of the involvement of other experts in the litigation.” NRC Panel, *supra* note 9, at 167.

12. See The Special Comm. on Empirical Data in Legal Decision Making, *Recommendations on Pretrial Proceedings in Cases with Voluminous Data*, reprinted in NRC Panel, *supra* note 9, app. F; David H. Kaye, *Improving Legal Statistics*, 24 Law & Soc’y Rev. 1255 (1990).

13. Drawing on “open science” practices, commentators have proposed adaptations of practices for “registration” of academic research studies. See Beerdsen, *supra* note 10; section titled “Recent Emphasis on the Limitations of *p*-values” below.

How Have the Data Been Collected?

The interpretation of data often depends on understanding “study design”—the plan for a statistical study and its implementation.¹⁴ Different designs are suited to answering different questions. Also, flaws in the data can undermine any statistical analysis, and data quality is often determined by study design.

In many cases, statistical studies are used to show causation. Do food additives cause cancer? Does capital punishment deter crime? Would additional disclosures in a securities prospectus cause investors to behave differently? The design of studies to investigate causation is the first topic of this section.¹⁵

Sample data can be used to describe a population. The population is the whole class of units that are of interest; the sample is the set of units chosen for detailed study. Inferences from the part to the whole are justified when the sample is representative. Sampling is the second topic of this section.

Finally, issues associated with the reliability and accuracy of collected data will be considered. Measurement error should be assessed and the likely impact of errors considered. Data quality is the third topic of this section.

All the sections concern the study of “variables.” In statistics, a variable is a characteristic of the units in a study. With a study of people, the unit of analysis is the person, and the variables describe people. Two such variables would be income (dollars per year) and educational level (years of schooling completed). With a study of school districts, the unit of analysis is the district. Typical variables include average family income of district residents and average test scores of students in the district.

Variables may be related to one another in various ways. Many studies examine whether a variable or group of variables, known as independent variables, are related to an outcome or dependent variable. For example, census data can be analyzed to determine whether people who complete more years of school tend to have higher incomes later in life. Educational level would be the independent variable, and annual income the dependent variable. In a study of smoking and lung cancer, the independent variable could be smoking (perhaps measured by the number of cigarettes smoked per day), and the dependent variable could mark the presence or absence of lung cancer.

14. For introductory treatments of data collection, see, for example, David Freedman et al., *Statistics* (4th ed. 2007); Darrell Huff, *How to Lie with Statistics* (1993); David S. Moore & William I. Notz, *Statistics: Concepts and Controversies* (10th ed. 2019); Hans Zeisel, *Say It with Figures* (6th ed. 1985); Hans Zeisel & David Kaye, *Prove It with Figures: Empirical Methods in Law and Litigation* (1997).

15. See also Steve C. Gold et al., *Reference Guide on Epidemiology*, “The Different Kinds of Epidemiologic Studies” section, in this manual.

Is the Study Designed to Investigate Causation?

Types of Studies

When causation is the issue, anecdotal evidence can be brought to bear. So can observational studies or controlled experiments. Anecdotal reports may be of value, but they are ordinarily more helpful in generating lines of inquiry than in proving causation. In medicine, evidence from clinical practice can be a good starting point for discovery of cause-and-effect relationships, but the anecdotal experience of practitioners is not definitive.¹⁶ Observational studies can establish that one factor is associated with another, but work is needed to bridge the gap between association and causation. Randomized controlled experiments are ideally suited for demonstrating causation.

Anecdotal evidence usually amounts to reports that events of one kind are followed by events of another kind. Typically, the reports are not even sufficient to show association, because there is no comparison group. For example, some children who live near power lines develop leukemia. Does exposure to electrical and magnetic fields cause this disease? The anecdotal evidence is not compelling because leukemia also occurs among children without exposure.¹⁷ It is necessary to compare disease rates among those who are exposed and those who are not. If exposure causes the disease, the rate should be higher among the exposed and lower among the unexposed. That would be association.

16. Consequently, many courts have suggested that attempts to infer causation from anecdotal reports are inadmissible as unsound methodology under *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 509 U.S. 579 (1993). See, e.g., *Miller v. Pfizer, Inc.*, 356 F.3d 1326, 1331 (10th Cir. 2004) (affirming the district court's exclusion in part because "placing substantial emphasis on a few challenge-dechallenge-rechallenge studies and case reports is not a generally accepted methodology"); *Hendrix ex rel. G.P. v. Evenflo Co., Inc.*, 609 F.3d 1183 (11th Cir. 2010) ("Case studies and clinical experience, used alone and not merely to bolster other evidence, are . . . insufficient to show general causation."); cf. *Matrixx Initiatives, Inc. v. Siracusano*, 563 U.S. 27, 44 (2011) (concluding that adverse-event reports combined with other information could be of concern to a reasonable investor and therefore subject to a requirement of disclosure under SEC Rule 10b-5, but stating that "the mere existence of reports of adverse events . . . says nothing in and of itself about whether the drug is causing the adverse events"). Other courts are more open to "differential diagnoses" based primarily on timing. E.g., *Best v. Lowe's Home Ctrs., Inc.*, 563 F.3d 171 (6th Cir. 2009) (reversing the exclusion of a physician's opinion that exposure to propenyl chloride caused a man to lose his sense of smell because of the timing in this one case and the physician's inability to attribute the change to something else).

17. See Nat'l Research Council, *Possible Health Effects of Exposure to Residential Electric and Magnetic Fields* (1997). There are problems in measuring exposure to electromagnetic fields, and results are inconsistent from one study to another. For such reasons, the epidemiologic evidence for an effect on health is inconclusive. Nat'l Cancer Inst., *Electromagnetic Fields and Cancer*, May 30, 2022 ("Studies have examined associations of these cancers with living near power lines, with magnetic fields in the home, and with exposure of parents to high levels of magnetic fields in the workplace. No consistent evidence for an association between any source of non-ionizing EMF and cancer has been found.").

The next issue is crucial: Exposed and unexposed people may differ in ways other than the exposure they have experienced. For example, children who live near power lines could come from poorer families and be more at risk from other environmental hazards. Such differences can create the appearance of a cause-and-effect relationship. Other differences can mask a real relationship. Cause-and-effect relationships often are quite subtle, and carefully designed studies are needed to draw valid conclusions.

An epidemiological classic makes the point. At one time, it was thought that lung cancer was caused by fumes from tarring the roads, because many lung cancer patients lived near roads that recently had been tarred. This is anecdotal evidence. But the argument is incomplete. For one thing, most people—whether exposed to asphalt fumes or unexposed—did not develop lung cancer. A comparison of rates was needed. Epidemiologists found that exposed persons and unexposed persons suffered from lung cancer at similar rates: Tar was probably not the causal agent. Exposure to cigarette smoke, however, turned out to be strongly associated with lung cancer. This study, in combination with later ones, made a compelling case that smoking cigarettes is the main cause of lung cancer.¹⁸

A good study design compares outcomes for subjects who are exposed to some factor (the treatment group) with outcomes for other subjects who are not exposed (the control group). With comparison groups, there is another important distinction—between controlled experiments and observational studies. In a controlled experiment, the investigators decide which subjects will be exposed and which subjects will be in the control group. In observational studies, the researchers do not determine which subjects are exposed; often, the subjects themselves choose their exposures. Because of self-selection or for other reasons, the treatment and control groups are likely to differ with respect to influential factors other than the ones of primary interest. These other factors are called lurking variables or confounding variables.¹⁹ A confounding variable can be correlated with the independent variable and act causally on the dependent variable. When

18. Richard Doll & A. Bradford Hill, *A Study of the Aetiology of Carcinoma of the Lung*, 2 Brit. Med. J. 1271 (1952), <https://doi.org/10.1136/bmj.2.4797.1271>. This was a matched case-control study. Cohort studies soon followed. See Gold et al., *supra* note 15. For a review of the evidence on causation, see 38 Int'l Agency for Research on Cancer (IARC), World Health Org., IARC Monographs on the Evaluation of the Carcinogenic Risk of Chemicals to Humans: Tobacco Smoking (1986) (updated 100E, A Review of Human Carcinogens: Personal Habits and Indoor Combustions 43–211 (2012)).

19. Epidemiologists sometimes limit “confounding” to “a bias due to the existence of a common cause of exposure and outcome” and define “selection bias” as “bias by [selecting units for study based] on common effects of otherwise unrelated variables.” Stephen R. Cole et al., *Illustrating Bias Due to Conditioning on a Collider*, 39 Int'l J. Epidemiology 417, 420 (2010), <https://doi.org/10.1093/ije/dyp334>. The distinction can be helpful in spotting different threats to causal inference, but this section uses the terms more loosely.

that happens, the confounder—not the independent variable—could be responsible for differences seen on the dependent variable. With the health effects of power lines, family background is a possible confounder; so is exposure to other hazards. Many confounders have been proposed to explain the association between smoking and lung cancer, but careful epidemiological studies have ruled them out, one after the other.

Confounding remains a problem even for the best observational research. For example, women with herpes are more likely to develop cervical cancer than other women. Some investigators concluded that herpes caused cancer: In other words, they thought the association was causal. Later research showed that the primary cause of cervical cancer was human papilloma virus (HPV). Herpes was a marker of sexual activity. Women who had multiple sexual partners were more likely to be exposed not only to herpes but also to HPV. The association between herpes and cervical cancer was due to other variables.²⁰

Randomized Controlled Experiments

In randomized controlled experiments, investigators assign subjects to treatment or control groups at random. The groups are therefore likely to be comparable, except for the treatment. This minimizes the role of confounding. Minor imbalances will remain, owing to the play of random chance; the likely effect on study results can be assessed by statistical techniques.²¹ The bottom line is that causal inferences based on well-executed randomized experiments are generally more secure than inferences based on well-executed observational studies.

The following example should help bring the discussion together. Today, we know that taking aspirin helps prevent heart attacks. But initially, there was some controversy. People who take aspirin rarely have heart attacks. This is anecdotal evidence for a protective effect, but it proves almost nothing. After all, few people have frequent heart attacks, whether or not they take aspirin regularly. A good study compares heart-attack rates for two groups: people who take aspirin (the treatment group) and people who do not (the controls). An observational study would be easy to do, but in such a study the aspirin-takers are likely to be different from the controls. Indeed, they are likely to be sicker—that is why they are taking aspirin. The study would be biased against finding a protective effect.

20. For additional examples and further discussion, see David A. Freedman, *From Association to Causation: Some Remarks on the History of Statistics*, 14 Stat. Sci. 243 (1999). Some studies find that herpes is a “cofactor,” which increases risk among women who are also exposed to HPV. Only certain strains of HPV are carcinogenic.

21. Randomization of subjects to treatment or control groups puts statistical tests of significance on a secure footing. See section titled “What Inferences Can Be Drawn from the Data?” below.

Randomized experiments are harder to do, but they provide better evidence.²² The experiments demonstrate the protective effect.²³

In summary, data from a treatment group without a control group generally reveal very little and can be misleading. Comparisons are essential. If subjects are assigned to treatment and control groups at random, a difference in the outcomes between the two groups can usually be accepted, within the limits of statistical error,²⁴ as a good measure of the treatment effect. However, if the groups are created in any other way, differences that existed before treatment may contribute to differences in the outcomes or mask differences that otherwise would become manifest. Observational studies succeed to the extent that the treatment and control groups are comparable—apart from the treatment.

Observational Studies

The bulk of the statistical studies seen in court are observational, not experimental. Take the question of whether capital punishment deters murder. To conduct a randomized controlled experiment, people would need to be assigned randomly to a treatment group or a control group. People in the treatment group would know they were subject to the death penalty for murder; the controls would know that they were exempt. Conducting such an experiment is not possible.

Many studies of the deterrent effect of the death penalty have been conducted, all observational, and some have attracted judicial attention. Researchers have catalogued differences in the incidence of murder in states with and without the death penalty and have analyzed changes in homicide rates and execution rates over the years. When reporting on such observational studies, investigators may speak of “control groups” (e.g., the states without capital punishment) or claim they are “controlling for” confounding variables by statistical methods such as multiple regression.²⁵ However, association is not causation. The

22. One important feature of a clinical trial is “blinding” to prevent unconscious bias. In a “double blind” trial, neither the patients nor the clinicians ascertaining the outcomes know who received the treatment and who did not. This prevents the expectations of the subjects and the experimenters from systematically affecting the observed outcomes in one group relative to the other.

23. But randomized experiments also show that aspirin can cause internal bleeding, raising the practical questions of whether and when a low-dose regimen is more beneficial than harmful. See, e.g., U.S. Preventive Services Task Force, *Aspirin Use to Prevent Cardiovascular Disease: Preventive Medication*, Apr. 26, 2022, <https://perma.cc/C8V9-WMYS>. In other instances, experiments have contradicted strongly held beliefs. E.g., Eric A. Klein et al., *Vitamin E and the Risk of Prostate Cancer: Results of the Selenium and Vitamin E Cancer Prevention Trial (SELECT)*, 306 JAMA 1549 (2011), <https://doi.org/10.1001/jama.2011.1437>.

24. See section titled “What Inferences Can Be Drawn from the Data?” below.

25. Multiple regression is described in Daniel L. Rubinfeld & David Card, *Reference Guide on Multiple Regression and Advanced Statistical Models*, in this manual. On the limits of regression to cope

causal inferences that can be drawn from analysis of observational data—no matter how complex the statistical technique—usually rest on a foundation that is less secure than that provided by randomized controlled experiments.

When an external change in circumstances naturally creates a treatment and a control group in a manner that is comparable to random assignment by the researchers, the study may be called a “natural experiment” or a “quasi-experiment.”²⁶ A celebrated example comes from Dr. John Snow’s investigation of cholera and sewage in the mid-1800s. The residents of an area in London received water from two companies with overlapping water mains drawing from a part of the Thames River that was heavily polluted with raw sewage. Then, one company moved its water intake upstream, to a less polluted spot. Snow showed that during a London cholera epidemic soon afterwards, the death rates from cholera in homes with the less polluted water was less than one-eighth that for the more polluted homes—and less than that for the rest of London.²⁷ Snow argued that “no experiment could have been devised which would more thoroughly test the effect of water supply on the progress of cholera than this”²⁸ In modern terminology, the argument is that even though the study was observational, the *seemingly* random assignment of homes to the two intake points protects against confounding and bias.

Observational studies can be very useful even when assignments to the groups being compared do not resemble randomization imposed by an experimenter. For example, there is strong observational evidence that smoking causes lung cancer (see section titled “Types of Studies” above). Generally, observational studies provide good evidence in the following circumstances:

- The association is seen in studies with different designs, on different kinds of subjects, and done by different research groups.²⁹ That reduces the chance that the association is due to a defect in one type of study, a peculiarity in one group of subjects, or the idiosyncrasies of one research group.

with lurking variables, see Richard A. Berk, *Regression Analysis: A Constructive Critique* (2004); Richard Berk, *What You Can and Can’t Properly Do with Regression*, 26 J. Quantitative Criminology 481 (2010), <https://doi.org/10.1007/s10940-010-9116-4>; David A. Freedman, *Statistical Models: Theory and Practice* (2005).

26. See, e.g., Frank de Vocht et al., *Conceptualising Natural and Quasi Experiments in Public Health*, 21 BMC Med. Research Methodology 32 (2021), <https://doi.org/10.1186/s12874-021-01224-x>.

27. John Snow, *On the Mode of Transmission of Cholera* 55–98 (2d ed. 1855).

28. *Id.* at 46.

29. For example, case-control studies are designed one way and cohort studies another, with many variations. See, e.g., David D. Celentano & Moyses Szklo, *Gordis Epidemiology* (6th ed. 2020); *supra* note 18.

- The association holds when effects of confounding variables are taken into account by appropriate methods, for example, comparing smaller groups that are relatively homogeneous with respect to the confounders.³⁰
- There is a plausible explanation for the effect of the independent variable; alternative explanations in terms of confounding should be less plausible than the proposed causal link.³¹ Thus, evidence for the causal link does not depend on observed associations alone.

Observational studies can produce legitimate disagreement among experts, and there is no mechanical procedure for resolving such differences of opinion. In the end, deciding whether associations are causal typically is not a matter of statistics alone, but also rests on scientific judgment.³²

There are, however, some basic questions to ask when appraising causal inferences based on empirical studies:

- Was there a control group? Unless comparisons can be made, the study has little to say about causation.
- If there was a control group, how were subjects assigned to treatment or control: through a process under the control of the investigator (a controlled experiment) or through a process outside the control of the investigator (an observational study)?
- If the study was a controlled experiment, was the assignment made using a chance mechanism (randomization), or did it depend on the judgment of the investigator?

If the data came from an observational study or a nonrandomized controlled experiment,

- How did the subjects come to be in treatment or in control groups?
- Are the treatment and control groups comparable?

30. The idea is to control for the influence of a confounder by stratification—making comparisons separately within groups for which the confounding variable is nearly constant and therefore has little influence over the variables of primary interest. For example, smokers are more likely to get lung cancer than nonsmokers. Age, gender, social class, and region of residence are all confounders, but controlling for such variables does not materially change the relationship between smoking and cancer rates.

31. A. Bradford Hill, *The Environment and Disease: Association or Causation?*, 58 *Proc. Royal Soc'y Med.* 295 (1965); Alfred S. Evans, *Causation and Disease: A Chronological Journey* 187 (1993).

32. At best, statistical analysis can help address the question of how impactful an unknown confounding variable would have to be to vitiate an inference of causation. See Jerome Cornfield et al., *Smoking and Lung Cancer: Recent Evidence and a Discussion of Some Questions*, 22 *J. Nat'l Cancer Inst.* 173 (1959), <https://doi.org/10.1093/jnci/22.1.173>; Tyler J. VanderWeele, *Are Greenland, Ioannidis and Poole Opposed to the Cornfield Conditions? A Defence of the E-value*, 51 *Int'l J. Epidemiology* 364 (2022), <https://doi.org/10.1093/ije/dyab218>.

- If not, what adjustments were made to address confounding?
- Were the adjustments sensible and sufficient?

Generalizing the Results

In considering what conclusions can be drawn from studies, it is helpful to distinguish between *internal* and *external* validity. Internal validity concerns the specifics of a particular study: Threats to internal validity include confounding and chance differences between treatment and control groups. External validity concerns using a particular study or set of studies to reach more general conclusions. A careful randomized controlled experiment on a large but unrepresentative group of subjects will have high internal validity but low external validity.

Any study must be conducted on certain subjects, at certain times and places, and using certain treatments. To extrapolate from the conditions of a study to more general conditions raises questions of external validity. For example, studies suggest that definitions of insanity given to jurors influence decisions in cases of incest. Would the definitions have a similar effect in cases of murder? Other studies indicate that recidivism rates for ex-convicts are not affected by providing them with temporary financial support after release. Would similar results be obtained if conditions in the labor market were different?

Confidence in the appropriateness of an extrapolation cannot come from the experiment itself. It comes from knowledge about outside factors that would or would not affect the outcome. Such judgments are easiest in the physical and life sciences, but even here, there are problems. For example, it may be difficult to infer human responses to substances that affect animals. First, there are often inconsistencies across test species. A chemical may be carcinogenic in mice but not in rats. Extrapolation from rodents to humans is even more problematic. Second, to get measurable effects in animal experiments, chemicals are administered at very high doses. Results are extrapolated—using mathematical models—to the very low doses of concern in humans. However, there are many dose–response models to use and few grounds for choosing among them. Generally, different models produce radically different estimates of the “virtually safe dose” in humans.³³

33. David A. Freedman & Hans Zeisel, *From Mouse to Man: The Quantitative Assessment of Cancer Risks*, 3 Stat. Sci. 3 (1988), <https://doi.org/10.1214/ss/1177012993>; Lorenz R. Rhomberg et al., *Linear Low-Dose Extrapolation for Noncancer Health Effects Is the Exception, Not the Rule*, 41 Critical Reviews Toxicology 1 (2011), <https://doi.org/10.3109/10408444.22010.536524>. For these reasons, many experts—and some courts in toxic tort cases—have concluded that evidence from animal experiments is generally insufficient by itself to establish causation. Likewise, extrapolation from animals to humans in testing for drug efficacy and safety has been the subject of extensive discussion. Johnique T. Atkins et al., *Pre-clinical Animal Models Are Poor Predictors of Human Toxicities in Phase 1 Oncology Clinical Trials*, 123 Brit. J. Cancer 1496 (2020), <https://doi.org/10.1038/s41416-020-01033-x>; Brian R. Berridge, *Animal Study Translation: The Other Reproducibility Challenge*,

Sometimes several studies, each having different limitations, all point in the same direction. This combination is why most experts believe that smoking causes lung cancer and many other diseases. So, too, a variety of studies indicate that jurors who approve of the death penalty are more likely to convict in a capital case.³⁴ Convergent results support the validity of generalizations.

Descriptive Surveys and Censuses

We now turn to a second topic—choosing units for study. A census tries to measure some characteristic of every unit in a population. This is often impractical. Then investigators use sample surveys, which measure characteristics for only part of a population. The accuracy of the information collected in a census or survey depends on how the units are selected for study and how the measurements are made.³⁵

What Method Is Used to Select the Units?

By definition, a census seeks to measure some characteristic of every unit in a whole population. It may fall short of this goal; in which case one must ask whether the missing data are likely to differ in some systematic way from the data that are collected.³⁶ The methodological framework of a scientific survey is different. With probability methods, a sampling frame (i.e., an explicit list of units in the

62 Int'l Lab'y Animal Rsch. J. 1 (2021), <https://doi.org/10.1093/ilar/ilac005>; John P. A. Ioannidis, *Extrapolating from Animals to Humans*, 12 Sci. Translational Med. 151 (2012), <https://doi.org/10.1126/scitranslmed.3004631>; Pandora Pound & Merel Ritskes-Hoiting, *Is It Possible to Overcome Issues of External Validity in Preclinical Animal Research? Why Most Animal Models Are Bound to Fail*, 16 J. Translational Med. 304 (2018), <https://doi.org/10.1186/s12967-018-1678-1>.

34. Phoebe C. Ellsworth, *Some Steps Between Attitudes and Verdicts*, in Inside the Juror 42, 46 (Reid Hastie ed., 1993). Nonetheless, in *Lockhart v. McCree*, 476 U.S. 162 (1986), the Supreme Court held that the exclusion of opponents of the death penalty in the guilt phase of a capital trial does not violate the constitutional requirement of an impartial jury.

35. See Shari Seidman Diamond et al., *Reference Guide on Survey Research*, “Population Definition and Sampling” section, in this manual.

36. The U.S. Decennial Census does not count everyone that it should, and it counts some people who should not be counted. There is evidence that net undercount is greater in some demographic groups than others. Supplemental studies may enable statisticians to adjust for errors and omissions. See Elizabeth Marra & Timothy Kennel, U.S. Census Bureau, PES20-J-01, *Source and Accuracy of the 2020 Post-Enumeration Survey Person Estimates: 2020 Post-Enumeration Survey Methodology Report* (2022). However, the adjustments rest on uncertain assumptions. See Lawrence D. Brown et al., *Statistical Controversies in Census 2000*, 39 Jurimetrics J. 347 (1999); David A. Freedman & Kenneth W. Wachter, *Methods for Census 2000 and Statistical Adjustments*, in *Social Science Methodology* 232 (Steven Turner & William Outhwaite eds., 2007) (reviewing technical issues and litigation surrounding census adjustment in 1990 and 2000).

population) is created. Individual units then are selected by an objective, well-defined, chance procedure, and measurements are made on the sampled units.

A sampling frame

To illustrate a sampling frame, suppose that a defendant in a criminal case seeks a change of venue. According to the defendant, popular opinion is so adverse that it would be difficult to impanel an unbiased jury. To prove the state of popular opinion, the defendant commissions a survey. The relevant population consists of everyone in the jurisdiction who might be called for jury duty. The sampling frame is the list of all potential jurors, which is maintained by court officials and is made available to the defendant. In this hypothetical case, the fit between the sampling frame and the population would be excellent.

In other situations, the sampling frame is more problematic. In an obscenity case, for example, the defendant can offer a survey of community standards.³⁷ The population comprises all adults in the legally relevant district, but obtaining a full list of such people may not be possible. Suppose the survey is done by telephone, but cell phones are excluded from the sampling frame. Suppose too that cell phone users, as a group, hold different opinions from landline users. Then the poll is unlikely to reflect the opinions of the cell phone users, no matter how many individuals are sampled and no matter how carefully the interviewing is done.³⁸

Selection bias

Many surveys do not use probability methods. In commercial disputes involving trademarks or advertising, the population of all potential purchasers of a product is hard to identify. Pollsters may resort to an easily accessible subgroup of the population—for example, shoppers in a mall. Such convenience samples may be

37. On the admissibility of such polls, see *State v. Midwest Pride IV, Inc.*, 721 N.E.2d 458 (Ohio Ct. App. 1998) (holding one such poll to have been properly excluded, and collecting cases from other jurisdictions); Admissibility of Evidence of Public-Opinion Polls or Surveys in Obscenity Prosecutions on Issue Whether Materials in Question Are Obscene, 59 A.L.R.5th 749.

38. Survey researchers may turn to other methods to reach cell phone users with area codes from the jurisdiction's geographic locale. Kyle McGeeney & Courtney Kennedy, Pew Research Center, *Advances in Telephone Survey Sampling* (2015), <https://perma.cc/D4MY-2L93>. However, the cell phone sampling frame will omit people who have moved into the jurisdiction with cell phone numbers from other areas. Again, the mismatch between the sampling frame and the relevant population could bias the results. People who move from county-to-county or state-to-state tend to be younger, more likely to be minority, to be male, and to have lower incomes. See Carol Pierannunzi et al., *Sample and Respondent Provided County Comparisons Among Cellular Respondents Using Rate Center Assignments*, 12 Survey Practice 1 (2019), <https://doi.org/10.29115/SP-2019-004>.

biased by the interviewer's discretion in deciding whom to approach—a form of selection bias—and the refusal of some of those approached to participate—nonresponse bias (see section titled “Of the Units Selected, Which Provide Measurements?” below). Selection bias is acute when constituents write their representatives, listeners call into radio talk shows, interest groups collect information from their members, individuals complete available online surveys, or attorneys choose cases for trial.³⁹

A well-known example of selection bias is the 1936 *Literary Digest* poll. After successfully predicting the winner of every U.S. presidential election since 1916, the *Digest* used the replies from 2.4 million respondents to predict that Alf Landon would win the popular vote, 57% to 43%. In fact, Franklin Roosevelt won by a landslide vote of 62% to 38%.⁴⁰ The *Digest* was so far off, in part, because it chose names from telephone books, rosters of clubs and associations, city directories, lists of registered voters, and mail order listings.⁴¹ In 1936, when only one household in four had a telephone, the people whose names appeared on such lists tended to be more affluent. Lists that overrepresented the affluent had worked well in earlier elections, when rich and poor voted along similar lines, but the bias in the sampling frame proved fatal when the Great Depression made economics a salient consideration for voters.

There are procedures that attempt to correct for selection bias. In quota sampling, for example, the interviewer is instructed to interview so many women, so many older people, so many ethnic minorities, and the like. But quotas still leave discretion to the interviewers in selecting members of each demographic group and therefore do not solve the problem of selection bias.⁴²

Probability methods are designed to avoid selection bias. Once the population is reduced to a sampling frame, the units to be measured are selected by a lottery that gives each unit in the sampling frame a known, nonzero probability of being chosen.⁴³ Such procedures are used to select individuals for jury duty.

39. *In re Chevron U.S.A., Inc.*, 109 F.3d 1016, 1020 (5th Cir. 1997) (although random sampling of 30 cases to resolve common issues or to ascertain damages in 3,000 claims arising from Chevron's allegedly improper disposal of hazardous substances would have been acceptable, having the opposing parties select 15 cases each was not, because those were “not cases calculated to represent the group of 3000 claimants”); *In re Countrywide Fin. Corp. Mortgage-Backed Sec. Litig.*, 984 F. Supp. 2d 1021, 1039 (C.D. Cal. 2013) (“The [sample's disproportionate] reliance on loans supporting certificates selected for litigation strikes the Court as a clear example of selection bias . . .”). See *infra* note 44 (on sampling cases or claims from a large set of similar cases or claims).

40. See Freedman et al., *supra* note 14, at 334–35.

41. *Id.* at 335, A–20 n.6.

42. See *id.* at 337–39.

43. Many types of probability sampling have been developed. In simple random sampling, units are drawn at random without replacement. In particular, each unit has the same probability of being chosen for the sample. *Id.* at 339–41. More complicated methods, such as stratified sampling and cluster sampling, give greater selection probabilities to some types of units than others, which has advantages in certain applications. In systematic sampling, every *n*th unit (for example, every

They also have been proposed or used to choose “bellwether” cases for representative trials to resolve issues in a large group of similar cases.⁴⁴

Of the Units Selected, Which Provide Measurements?

Probability sampling ensures that within the limits of chance, the sample will be representative of the sampling frame. But will all these units be measured? When documents are sampled for audit, they can all be examined, at least in principle. Human beings are less easily managed, and some will refuse to cooperate. In the 1936 *Literary Digest* election poll, only 24% of the 10 million people who received questionnaires returned them. Most of the respondents probably had strong views on the candidates and objected to President Roosevelt’s economic program. This self-selection is likely to have biased the poll.⁴⁵

Surveys should therefore report nonresponse rates. A large nonresponse rate warns of potential bias.⁴⁶ Supplemental studies may establish that nonrespondents are similar to respondents with respect to characteristics of interest, but

fifth, tenth, or hundredth unit) in the sampling frame is selected. If the units are not in any special order, then systematic sampling is often comparable to simple random sampling.

44. See *Scottsdale Mem’l Health Sys., Inc. v. Maricopa Cnty.*, 228 P.3d 117, 131–35 (Ariz. Ct. App. 2010) (reviewing major cases and approving in principle of random sampling but concluding that the record failed to demonstrate the adequacy of the statistical sampling plan for resolving 40,000 consolidated cases); David H. Kaye & David A. Freedman, *Statistical Proof*, in 1 *Modern Scientific Evidence: The Law and Science of Expert Testimony* § 5:16, at 398–406 (David L. Faigman et al. eds., 2022–2023) (discussing both legal and statistical issues arising in these cases); David H. Kaye et al., *The New Wigmore: A Treatise on Evidence: Expert Evidence* § 12.10.3(b) (Cumulative Supp. to 2d ed., 2020) (same).

45. Maurice C. Bryson, *The Literary Digest Poll: Making of a Statistical Myth*, 30 *Am. Statistician* 184 (1976); Freedman et al., *supra* note 14, at 335–36.

46. For discussions of the admissibility of surveys with extremely low response rates, see *In re ConAgra Foods, Inc.*, 90 F. Supp. 3d 919 (C.D. Cal. 2015) (excluding survey with many defects, including a 95% nonresponse rate); *United States v. H & R Block, Inc.*, 831 F. Supp. 2d 27, 34 (D.D.C. 2011) (98% nonresponse rate does not preclude admission when a judge is the factfinder). On non-response rates in studies not undertaken for litigation, see Clearinghouse for Military Family Readiness, Penn. State Univ., *Survey Response Rates: Rapid Literature Review* (2016), available at <https://perma.cc/8Y9Z-PTKH>.

In *United States v. Gometz*, 730 F.2d 475, 478 (7th Cir. 1984) (en banc), the Seventh Circuit recognized that “a low rate of response to juror questionnaires could lead to the underrepresentation of a group that is entitled to be represented on the qualified jury wheel.” Nonetheless, the court held that under the Jury Selection and Service Act of 1968, 28 U.S.C. §§ 1861–1878 (1988), the clerk did not abuse his discretion by failing to take steps to increase a response rate of 30%. According to the court, “Congress wanted to make it possible for all qualified persons to serve on juries, which is different from forcing all qualified persons to be available for jury service.” *Gometz*, 730 F.2d at 480. Although it might “be a good thing to follow up on persons who do not respond to a jury questionnaire,” the court concluded that Congress “was not concerned with anything so esoteric as nonresponse bias.” *Id.* at 479, 482; *cf. In re United States*, 426 F.3d 1 (1st Cir. 2005)

even when demographic characteristics of the sample match those of the population, caution is indicated.⁴⁷

In short, a good survey defines an appropriate population, uses a probability method for selecting the sample, has a high response rate, and gathers accurate information on the sample units. When these goals are met, the sample tends to be representative of the population, and data from the sample can be extrapolated to describe the characteristics of the population. Of course, surveys may be useful even if they fail to meet these criteria. But then, additional arguments are needed to justify the inferences.

Individual Measurements

Is the Measurement Process Reliable?

Reliability and validity are two aspects of accuracy in measurement. In statistics, reliability refers to reproducibility of results. A reliable measuring instrument returns consistent measurements. A scale, for example, is perfectly reliable if it always reports the same weight for the same unchanged object. It may not be accurate—it may always report a weight that is too high or one that is too low—but the perfectly reliable scale always reports the same weight for the same object. Its errors, if any, are systematic: They tend to point in the same direction.

Courts often use “reliable” to mean “that which can be relied on” for some purpose, such as establishing probable cause through a reliable informant or as part of an argument for admitting hearsay statements.⁴⁸ Thus, in *Daubert v. Merrell Dow Pharmaceuticals*, the Court distinguished “evidentiary reliability” from reliability in the statistical sense of giving consistent results.⁴⁹ This statistical reliability is a component of the broader evidentiary reliability required of scientific evidence. It can be ascertained by measuring the same quantity several times; the measurements must be made independently to avoid bias. Given independence, the correlation coefficient (see section titled “Correlation Coefficients” below) between repeated measurements can be used as a measure of reliability. This is sometimes called a test-retest correlation or a reliability coefficient. But administering the same test twice to the same group of people may be impractical. And

(reaching the same result with respect to underrepresentation of African Americans resulting in part from nonresponse bias).

47. See David Streitfeld, *Shere Hite and the Trouble with Numbers*, 1 *Chance* 26 (1988); Chamont Wang, *Sense and Nonsense of Statistical Inference: Controversy, Misuse, and Subtlety* 174–76 (1992).

48. *E.g.*, *United States v. Moore*, 824 F.3d 620 (7th Cir. 2016) (“The purpose of Rule 807 is to make sure that reliable, material hearsay evidence is admitted, regardless of whether it fits neatly into one of the exceptions enumerated in the Rules of Evidence.”); Fed. R. Evid. 803(18)(B) (to be admissible hearsay, a learned treatise must be a “reliable authority”).

49. 509 U.S. 579, 590 n.9 (1993).

even if repeated testing is practical, it may be statistically inadvisable, because subjects may learn something from the first round of testing that affects their scores on the second round. Such “practice effects” are likely to compromise the independence of the two measurements, and independence is needed to estimate reliability. Statisticians therefore use internal evidence from the test itself. For example, a strong correlation between scores on the first half of the test and scores on the second half is evidence of reliability.⁵⁰

The Supreme Court was faced with the imperfect reliability of a test for IQ scores in *Hall v. Florida*.⁵¹ The Court listed various factors contributing to the variability in an individual’s test scores, including “health; practice from earlier tests; the environment or location of the test; the examiner’s demeanor; the subjective judgment involved in scoring certain questions on the exam; and simple lucky guessing.”⁵² Having previously held that intellectually disabled offenders could not be punished by execution, the Court held that a state had to account for the potential variability in an individual’s IQ score in setting a number above which no one could be deemed intellectually disabled. The particular rule the Court adopted turned on one version of a statistic known as the standard error.⁵³

A simpler courtroom example comes from DNA identification. An early method of identification required laboratories to determine the lengths of fragments of DNA. By making independent repeated measurements of the same fragments, laboratories determined the likelihood that two measurements differed by specified amounts.⁵⁴ Such results were needed to decide whether a discrepancy between a crime sample and a suspect sample was sufficient to exclude the suspect.⁵⁵

Coding of data also can affect reliability. In many studies, descriptive information is obtained on the subjects. For statistical purposes, the information usually has to be reduced to numbers. The process of reducing information to numbers is called “coding,” and the reliability of the process should be evaluated. For example, in a meticulous study of death sentencing in Georgia, legally trained evaluators examined short summaries of cases and ranked them according to the

50. In practice, more refined measures of internal consistency are used to estimate a test’s reliability. See, e.g., Neal M. Kingston & Laura B. Kramer, *High Stakes Test Construction and Test Use*, in 1 Oxford Handbook of Quantitative Methods: Foundations 189, 201 (Todd D. Little ed., 2013).

51. 572 U.S. 701 (2014).

52. *Id.* at 713.

53. “Standard error” is the subject of the section titled “Is a Difference Statistically Significant?” below. The relationship between the reliability coefficient and different standard errors as well as the choice of a cut-off score that accounts for a given type of standard error is explained in David H. Kaye, *Deadly Statistics: Quantifying an “Unacceptable Risk” in Capital Punishment*, 16 Law, Probability & Risk 7 (2017) (proposing alternatives to the Court’s rule).

54. See Nat’l Research Council, *The Evaluation of Forensic DNA Evidence* 139–41 (1996).

55. *Id.*; Nat’l Research Council, *DNA Technology in Forensic Science* 61–62 (1992). Current methods are discussed in David H. Kaye, *Reference Guide on Human DNA Identification Evidence*, in this manual.

defendant's culpability.⁵⁶ Two different aspects of reliability should be considered. First, the “within-observer variability” of judgments should be small—the same evaluator should rate essentially identical cases in similar ways. Second, the “between-observer variability” should be small—different evaluators should rate the same cases in essentially the same way.

Metrologists (specialists in measurement science) similarly distinguish between “reproducibility” and “repeatability.” The difference lies in the degree to which the conditions for making measurements are similar. With a repeatable procedure, measurements by the same examiner using the same equipment at the same place and time should be close to one another. With a reproducible procedure, measurements from different examiners even at different places and times also should be similar.⁵⁷

Is the Measurement Process Valid?

Reliability is necessary but not sufficient to ensure accuracy. In addition to reliability, validity is needed. A valid measuring instrument measures what it is supposed to. Thus, a polygraph measures certain physiological variables, for example, pulse rate or blood pressure, in response to stimuli. The measurements may be reliable. Nonetheless, the polygraph is not valid as a lie detector unless the measurements are well correlated with lying.⁵⁸

When there is an established way of measuring a variable, a new measurement process can be validated by comparison with the established one. Breathalyzer readings can be validated against alcohol levels found in blood samples. LSAT or GRE scores used for law school admissions can be validated against grades earned in law school. A common measure of validity is the correlation coefficient between the predictor and the criterion (for example, test scores and later performance).⁵⁹

Employment discrimination cases illustrate some of the difficulties. Plaintiffs suing under Title VII of the Civil Rights Act may challenge an employment

56. David C. Baldus et al., *Equal Justice and the Death Penalty: A Legal and Empirical Analysis* 49–50 (1990).

57. Alan H. Dorfman & Richard Valliant, *A Re-Analysis of Repeatability and Reproducibility in the Ames-USDOE-FBI Study*, 9 Stat. & Public Pol'y 175 (2022), <https://doi.org/10.1080/2330443X.2022.2120137>; Hal S. Stern et al., *Reliability and Validity of Forensic Science Evidence*, Significance, Apr. 2019, at 21, 22–23.

58. See *United States v. Henderson*, 409 F.3d 1293, 1303 (11th Cir. 2005) (“while the physical responses recorded by a polygraph machine may be tested, ‘there is no available data to prove that those specific responses are attributable to lying’”); Nat'l Research Council, *The Polygraph and Lie Detection* (2003) (reviewing the scientific literature).

59. As the discussion of the correlation coefficient in the section titled “Correlation Coefficients” below indicates, the closer the coefficient is to 1, the greater the validity. For a review of data on test reliability and validity, see *Measuring Success: Testing, Grades, and the Future of College Admissions* (Jack Buckley et al. eds., 2018).

test that has a disparate impact on a protected group, and defendants may try to justify the use of a test as valid, reliable, and a business necessity.⁶⁰ For validation, the most appropriate criterion variable is clear enough: job performance. However, plaintiffs may then challenge the validity of performance ratings. Are they sufficiently related to the actual requirements of the job? Are they biased? As for reliability, plaintiffs would need to show that each measure (the test scores and the later performance ratings) provides consistent measurements.⁶¹

A further problem is that test-takers are likely to be a select group. The ones who get the jobs are even more highly selected. Generally, selection attenuates (weakens) the correlations because differences in performance within a narrow band of highly qualified applicants do not exhibit as much variation as would be expected if a wider range of applicants were selected.⁶² Statistical methods that correct for attenuation depend on assumptions about the nature of the test and the procedures used to select the test-takers; these assumptions may be open to challenge.⁶³

Measurements also can be made on a nominal scale, as when a chemist notes that litmus paper has turned red, or a criminalist declares that there is a “physical fit” between two fragments of glass. For binary (yes–no) classifications, validity (accuracy) of the classifier can be evaluated with experiments to estimate “sensitivity” and “specificity” (or corresponding error probabilities). Suppose that fire-arms examiners are given pairs of spent cartridge cases and are required to decide whether they come from the same gun or from two different guns. The experimenter, who has prepared the test pairs, knows the truth; the examiners do not. Results related to a small experiment are in Table 1.⁶⁴

60. See, e.g., *Ricci v. DeStefano*, 557 U.S. 557 (2009); *Washington v. Davis*, 426 U.S. 229, 252 (1976); *Albemarle Paper Co. v. Moody*, 422 U.S. 405, 430–32 (1975); *Griggs v. Duke Power Co.*, 401 U.S. 424 (1971); *Lanning v. S.E. Penn. Transp. Auth.*, 308 F.3d 286 (3d Cir. 2002).

61. See section titled “Is the Measurement Process Reliable?” above (discussing *Hall v. Florida*).

62. For an extreme example, consider a firm that provides personal tutoring for college-entrance examinations and hires only job applicants who had near-perfect scores themselves to be the tutors. It could well be that receiving a high score is associated with being an effective tutor. But if the firm hired no low-scoring applicants as tutors, it would be impossible to see that association in a study of the correlation between the scores of the exclusively high-scoring tutors and those of the students they tutor after being hired.

63. See Thad Dunning & David A. Freedman, *Modeling Selection Effects*, in *Social Science Methodology* 225 (Steven Turner & William Outhwaite eds., 2007); Howard Wainer & David Thissen, *True Score Theory: The Traditional Method*, in *Test Scoring* 23 (David Thissen & Howard Wainer eds., 2001).

64. The numbers are rounded-off versions of those in Table C1 of Heike Hofmann et al., *Treatment of Inconclusives in the AFTE Range of Conclusions*, 19 *Law, Probability & Risk* 317, 363 (2020), <https://doi.org/10.1093/lpr/mgab002> (deducing numbers for independent comparisons within a somewhat differently designed and harder to interpret 2003 experiment with eight FBI examiners). That there are zeros in the cells for false identifications and false eliminations does not mean that these outcomes cannot occur. See the “Other situations” subsection for confidence intervals below. Also, in

Table 1. Test Results for Cartridge-case Comparisons

	Same Source	Different Source
Reported Same Source	30	0
Reported Different Source	0	45

The examiners performed flawlessly in these 75 instances. If we call the same-source condition “positive” and the “different-source” condition “negative,” there were zero false-positive reports and zero false-negative reports. To put it another way, the examiners were 100% accurate in dealing with same-source pairs—they called 30 out of 30 such pairs positives, for an observed sensitivity of 1; likewise, they were 100% accurate in dealing with different-source pairs—they called 45 out of 45 such pairs different, for an observed specificity of 1. Whether the results of this small experiment are representative of what might be seen in a larger set of experiments, and whether the experiments would be representative of the outcomes in case work are further questions.

Are the Measurements Recorded Correctly?

Judging the adequacy of data collection involves an examination of the process by which measurements are taken. Are responses to interviews coded correctly? Do mistakes distort the results? How much data are missing? What was done to compensate for gaps in the data? These days, data are stored in computer files. Cross-checking the files against the original sources (e.g., paper records), at least on a sample basis, can be informative.

Data quality is a pervasive issue in litigation and in applied statistics more generally.⁶⁵ A programmer moves a file from one computer to another, and half the data disappear. The definitions of crucial variables are lost in the sands of time. Values get corrupted: Social Security numbers come to have eight digits instead of nine, and vehicle identification numbers fail the most elementary consistency checks. Everybody in the company, from the CEO to the rawest mailroom trainee,

experiments and in practice, firearms examiners often report comparisons as “inconclusive.” This complication is discussed in the section titled “Are the Categories Appropriate?” below.

65. *E.g.*, Bland–Collins v. Howard Univ., 19 F. Supp. 3d 252, 255 (D.D.C. 2014) (statistician who was fired after she discovered “over 5,000 errors in the coded data collected from structured interviews” in a National Science Foundation funded research project brought a whistleblower retaliation claim). Transcription errors put the FBI in the awkward position of using 33 incorrect allele frequencies (out of 1,100) in the software it distributed since 1999 to DNA laboratories for computing random-match probabilities. Tamyra R. Moretti et al., *Erratum*, 60 J. Forensic Sci. 1114 (2015), <https://doi.org/10.1111/1556-4029.12806>; Spencer S. Hsu, *FBI Notifies Crime Labs of Errors Used in DNA Match Calculations Since 1999*, Wash. Post, May 29, 2015.

turns out to have been hired on the same day. Many of the residential customers have last names that indicate commercial activity (e.g., “Happy Valley Farriers”). These problems seem humdrum by comparison with those of reliability and validity, but—unless caught in time—they can be fatal to statistical arguments.⁶⁶

What Does It Mean to Be Random?

In the law, a selection process sometimes is called “random,” provided that it does not exclude identifiable segments of the population. Statisticians use the term in a more rigorous and technical sense. For example, to choose one person at random from a population in the strict statistical sense, we would have to ensure that everybody in the population has the same probability of selection. With a randomized controlled experiment, subjects are assigned to treatment or control at random in the strict sense—by tossing coins, throwing dice, looking at tables of random numbers, or more commonly these days, by using a random number generator on a computer. The same rigorous definition applies to random sampling. Randomness in the technical sense provides assurance of unbiased estimates from a randomized controlled experiment or a probability sample. Randomness in the technical sense also justifies calculations of standard errors, confidence intervals, and *p*-values (see sections titled “What Inferences Can Be Drawn from the Data?” and “Correlation and Regression” below). Looser definitions of randomness are inadequate for statistical purposes.

How Have the Data Been Presented?

After data have been collected, they should be presented in a way that makes them intelligible and that helps reveal their implications. Data can be summarized with a few numbers or with graphical displays. However, the wrong summary can

66. See, e.g., *Malletier v. Dooney & Bourke, Inc.*, 525 F. Supp. 2d 558, 630 (S.D.N.Y. 2007) (coding errors contributed “to the cumulative effect of the methodological errors” that warranted exclusion of a consumer confusion survey); *EEOC v. Sears, Roebuck & Co.*, 628 F. Supp. 1264, 1304, 1305 (N.D. Ill. 1986) (finding that the EEOC “has made so many general coding errors that its data base does not fairly reflect the characteristics of applicants”), *aff’d*, 839 F.2d 302 (7th Cir. 1988); cf. *EEOC v. Freeman*, 961 F. Supp. 2d 783, 796 (D. Md. 2013) (excluding an EEOC analysis of an employer’s records purporting to show disparate impact of credit and criminal records checks of job applicants because the EEOC database was not a random sample but rather was “cherry-picked” and “[t]he mind-boggling number of errors contained in [its] database could alone render [the] conclusions worthless”).

mislead.⁶⁷ The section “Are Rates or Percentages Properly Interpreted?” below discusses rates or percentages and provides some cautionary examples of misleading summaries, indicating the kinds of questions that might be considered when summaries are presented in court. Percentages are often used to demonstrate statistical association, which is the topic of the section titled “Is an Appropriate Measure of Association Used?” The section titled “Does a Graph Portray Data Fairly?” considers graphical summaries of data, while the sections titled “Is an Appropriate Measure Used for the Center of a Distribution?” and “Is an Appropriate Measure of Variability Used?” discuss some of the basic descriptive statistics that are likely to be encountered in litigation, including the mean, median, and standard deviation.

Are Rates or Percentages Properly Interpreted?

How Big Is the Base of a Percentage?

Rates and percentages often provide effective summaries of data, but these statistics can be misinterpreted. A rate reports a comparison of one number against some other quantity, for example, the number of reported crimes per hundred thousand residents. A percentage reports a comparison between two numbers by putting them in terms of a common base (100). Expressing the ratio of the two numbers on a common base makes it easy to compare them.

One application of percentages is for reporting increases or decreases in a quantity or rate by describing the percent change compared to the initial or base amount. When the base is small, however, a small change in absolute terms can generate a large percentage gain or loss. (This could lead to newspaper headlines such as “Increase in Rate of Thefts Alarming,” even when the total number of thefts is small.⁶⁸) Conversely, a large base will make for small percentage increases. In these situations, actual numbers may be more revealing than percentages.

Have Appropriate Benchmarks Been Provided?

The selective presentation of numerical information is like quoting someone out of context. Is the fact that a particular actively managed fund of large-cap stocks boasted a return of 25% in 2021 indicative of outstanding management? Considering that the 500 large-cap stocks in “the benchmark S&P 500 notched a total

67. See generally Freedman et al., *supra* note 14; Huff, *supra* note 14; Moore & Notz, *supra* note 14; Zeisel, *supra* note 14.

68. Lyda Longa, *Increase in Thefts Alarming*, Daytona News-J. June 8, 2008 (reporting a 35% increase in armed robberies in Daytona Beach, Florida, in a 5-month period, but not indicating whether the number had gone up by 6 (from 17 to 23), by 300 (from 850 to 1,150), or by some other amount).

return . . . of 28.7% [that] year,” a growth rate of 25% is less indicative of unusual financial acumen than one might have thought.⁶⁹ In this example and many others, it is helpful to find a benchmark that puts the figures into perspective.⁷⁰

Have the Data Collection Procedures Changed?

Changes in the process of collecting data can create problems of interpretation. Statistics on crime provide many examples. The number of petty larcenies reported in Chicago more than doubled one year—not because of an abrupt crime wave, but because a new police commissioner introduced an improved reporting system.⁷¹ For a time, police officials in Washington, D.C., “demonstrated” the success of a law-and-order campaign by valuing stolen goods at \$49, just below the \$50 threshold then used for inclusion in the FBI’s Uniform Crime Reports.⁷² Allegations of manipulation in the reporting of crime from one time period to another are legion.⁷³ Almost all series of numbers that cover many years are affected by changes in definitions and collection methods. When a study includes such time-series data, it is useful to inquire about changes in the way data are collected or reported and to look for any sudden jumps, which may signal such changes.

69. Karen Langley, *Stock Pickers Watched the S&P 500 Pass Them By Again in 2021*, Wall St. J., Mar. 16, 2022 (reporting that “[t]he failure of stock pickers to beat the benchmark is nothing new: 2021 was the 12th consecutive year in which the majority of actively-managed funds of large-cap stocks watched the S&P 500 pass them by”).

70. The selection of the benchmark may merit scrutiny. Securities and Exchange Commission (SEC) Rule 33–698 requires mutual funds to display past returns alongside those of “an appropriate broad-based securities market index.” But the rule “does not prohibit funds from comparing their past returns to those of newly-chosen index(es).” Kevin Mullally & Andrea Rossi, *Moving the Goalposts? Mutual Fund Benchmark Changes and Performance Manipulation*, June 24, 2022, <https://perma.cc/6MFU-WYEN>. There is evidence that, to attract more investors, funds simply change their benchmark indexes to lower performing ones—including benchmarks that are not the best match for the funds’ actual investment strategies. *Id.*

71. James P. Levine et al., *Criminal Justice in America: Law in Action* 99 (1986) (referring to a change from 1959 to 1960).

72. David Seidman & Michael Couzens, *Getting the Crime Rate Down: Political Pressure and Crime Reporting*, 8 Law & Soc’y Rev. 457 (1974).

73. John A. Eterno & Eli B. Silverman, *The Crime Numbers Game: Management by Manipulation* (2012); Michael D. Maltz, *Missing UCR Data and Divergence of the NCVS and UCR Trends, in Understanding Crime Statistics: Revisiting the Divergence of the NCVS and UCR* 269, 280 (James P. Lynch & Lynn A. Addington eds., 2006), <https://doi.org/10.1017/CBO9780511618543.010> (citing newspaper reports in Boca Raton, Atlanta, New York, Philadelphia, Broward County (Florida), and St. Louis). Changes in reporting practices also can be improvements, but they can still distort comparisons to data collected before the changes. *E.g.*, Greg Moreau, *Statistics Canada, Police-Reported Crime Statistics in Canada* 2018, at 7 (2019), <https://perma.cc/AR6J-DFTF>.

Are the Categories Appropriate?

Misleading summaries also can be produced by the categories used for comparison. In *Philip Morris, Inc. v. Loew's Theatres, Inc.*,⁷⁴ and *R.J. Reynolds Tobacco Co. v. Loew's Theatres, Inc.*,⁷⁵ Philip Morris and R.J. Reynolds sought an injunction to stop the maker of Triumph low-tar cigarettes from running advertisements claiming that participants in a national taste test preferred Triumph to other brands. Plaintiffs alleged that claims that Triumph was a “national taste test winner” or Triumph “beats” other brands were false and misleading. An exhibit introduced by the defendant contained the data shown in Table 2.⁷⁶ Only 14% + 22% = 36% of the sample preferred Triumph to Merit, whereas 29% + 11% = 40% preferred Merit to Triumph. By selectively combining categories, however, the defendant attempted to create a different impression. Because 24% found the brands to be about the same, and 36% preferred Triumph, the defendant claimed that a clear majority (36% + 24% = 60%) found Triumph “as good [as] or better than Merit.”⁷⁷ The court resisted this chicanery, finding that defendant’s test results did not support the advertising claims.⁷⁸

Table 2. Data Used by a Defendant to Refute Plaintiff’s False Advertising Claims

	Triumph Much Better	Triumph Somewhat Better	Triumph About the Same	Triumph Somewhat Worse	Triumph Much Worse
Number	45	73	77	93	36
Percentage	14	22	24	29	11

There was a similar distortion in claims for the accuracy of a home pregnancy test. The manufacturer advertised the test as 99.5% accurate under laboratory conditions. The underlying data are summarized in Table 3.

Table 3. Home Pregnancy Test Results

	Actually Pregnant	Actually Not Pregnant
Test says pregnant	197	0
Test says not pregnant	1	2
Total	198	2

74. 511 F. Supp. 855 (S.D.N.Y. 1980).

75. 511 F. Supp. 867 (S.D.N.Y. 1980).

76. *Philip Morris*, 511 F. Supp. at 866.

77. *Id.*

78. *Id.* at 856–57.

The table does indicate that only one error occurred in 200 assessments, for 99.5% overall accuracy. But the table also shows that the test can make two types of errors: It can tell a pregnant woman that she is not pregnant (a false negative), and it can tell a woman who is not pregnant that she is (a false positive). The reported 99.5% accuracy rate conceals a crucial fact—the company had virtually no data with which to measure the rate of false positives.⁷⁹

The problem with combining categories into broader ones has surfaced in criminal cases as well. As noted in the section titled “Is the Measurement Process Valid?” above, criminalists examine pairs of items, such as spent cartridge cases, to decide whether they are associated with the same source. In firearms-toolmark matching, the traditional comparison between an item of known origin and the item whose origin is in question culminates in a report of “identification,” “inconclusive,” or “elimination.” In validity studies, toolmark examiners are given items to evaluate when the experimenter, but not the criminalist, knows whether the items are from the same source or from two different sources. Table 4 expands Table 1 with a row for “inconclusives” in the experiment:

Table 4. Test Results for Cartridge-case comparisons⁸⁰

	Same Source	Different Source
Reported Same Source	30	0
Reported Different Source	0	45
Reported Inconclusive	0	157

The new row reveals that the criminalists reached no definitive conclusion in most of the cases, and that all these “inconclusives” pertained to pairs of cartridge cases from two different guns. Adding in all these “inconclusives,” the overall proportion of correct decisions drops from 100% to only $(30 + 45) / (30 + 45 + 157) = 32\%$.

79. Only two women in the sample were not pregnant; the test gave correct results for both of them (a specificity of 100%). Although a false-positive rate of 0 is ideal, an estimate based on a sample of only two women is not. These data are reported in Arnold Barnett, *How Numbers Can Trick You*, Tech. Rev., Oct. 1994, at 38, 44–45. When data on test performance are sufficient to accurately estimate (1) the probability of a positive when the condition is present (sensitivity) and (2) the probability of a negative when it is not (specificity), the estimates can be combined to express the probative value of a test result in the form of a likelihood ratio. See Appendix section titled “What Is Bayes’ Rule?” below. Ideally, the test should have both high sensitivity and specificity. However, the combination of moderate sensitivity and high specificity also indicates that a positive finding strongly supports the conclusion that the condition is present. *Id.*

80. The numbers are similar to those in Hofmann et al., *supra* note 64, at 363 (tbl. C1). More recent and larger studies with different results also are noted there.

In one sense, this is a poor “correct decision rate.”⁸¹ It indicates that examiners are missing many opportunities to reach conclusions; perhaps, if they were more venturesome, they would produce more evidence correctly excluding or correctly implicating suspects. Or perhaps they would simply make more mistakes. In any event, as with the advertised accuracy rate of 99.5% for the home pregnancy test, the overall correct-decision rate masks a crucial fact—here, that the examiners in the study, when confronted with marks from different guns, never erred in attributing them to the same gun. The low false-positive rate (0/45 incorrect same-source decisions) is more pertinent to a same-source determination offered to implicate the defendant than is the aggregated correct-decision rate.⁸² However, observing relatively few false positives (or even none at all, as in this study) may not be convincing—particularly with small or unrepresentative samples—and better measures for the probative value of a positive test result are available.⁸³

What Comparisons Are Made?

Finally, there is the issue of which numbers to compare. Researchers sometimes choose among alternative comparisons. Why did they choose the one they did?

81. See *People v. Winfield*, No. 15 CR 14066–01 (Ill. Cir. Ct. Feb. 9, 2023) (aff. of Susan Vanderplas, Kori Khan, Heike Hofmann, Alicia Carriquiry, Jan. 3, 2022) (affidavit submitted in murder case characterizing the “correct source decision rates” in experiments comparable to the one described here as “abysmal”).

82. Some scientists and litigants maintain that, by definition, every inconclusive result is an error (and should be counted as such) because it does not express the true association (positive or negative) within each pair. *Id.* The opposing view is that “inconclusive” is more of a conservative decision to opt-out of the binary classification scheme shown in Table 1 than an inference to the true state of affairs. Under this view, forensic scientists and policy makers might wish to investigate whether examiners should be more willing to make inclusions or exclusions (or to describe the strength of the evidence on a more graduated scale). But a report of “I cannot tell” is neither an erroneous nor a correct statement about the true source. This understanding has led commentators to argue that the number of “inconclusives” does not belong in either the numerator or the denominator of the proportions used to assess the validity of the positive or negative findings of association. See, e.g., Hal R. Arkes & Jonathan J. Koehler, *Inconclusives and Error Rates in Forensic Science: A Signal Detection Theory Approach*, 20 *Law, Probability & Risk* 153 (2021), <https://doi.org/10.1093/lpr/mgac005>. In response, it has been said that most inconclusives in validity studies would be false inclusions in practice, which can render an error rate that does not incorporate inconclusives misleading. In short, a range of views currently exists regarding what to count as “errors” in validity studies when extrapolating to actual casework. See Hal R. Arkes & Jonathan J. Koehler, *Inconclusive Conclusions in Forensic Science: Rejoinders to Scurich, Morrison, Sinha and Gutierrez*, 21 *Law, Probability & Risk* 175 (2022), <https://doi.org/10.1093/lpr/mgad002>.

83. We discuss sampling error in false-positive proportions in the section titled “What Is the Standard Error?” below and describe a more complete measure of probative value in the Appendix section titled “What Is Bayes’ Rule?”

Would another comparison give a different view? A government agency, for example, may want to compare the amount of service now being given with that of earlier years—but what earlier year should be the baseline? If the first year of operation is used, a large percentage increase should be expected because of startup problems. If last year is used as the base, was it also part of the trend, or was it an unusually poor year? If the base year is not representative of other years, the percentage may not portray the trend fairly. No single question can be formulated to detect such distortions, but it may help to ask for the numbers from which the percentages were obtained; asking about the base can also be helpful.⁸⁴

Is an Appropriate Measure of Association Used?

Many cases involve statistical association. Does a test for employee promotion have an exclusionary effect that depends on race or sex? Does the incidence of murder vary with the rate of executions for convicted murderers? Do consumer purchases of a product depend on the presence or absence of a product warning? This section discusses tables and percentage-based statistics that are frequently presented to answer such questions.⁸⁵

Percentages often are used to describe the association between two variables. Suppose that a university is alleged to discriminate against women in admitting students, and that the university consists of only two colleges—engineering and business. The university admits 350 out of 800 male applicants; by comparison, it admits only 200 out of 600 female applicants. Such data commonly are displayed as in Table 5.⁸⁶

Table 5. Admissions by Sex

Decision	Male	Female	Total
Admit	350	200	550
Deny	450	400	850
Total	800	600	1,400

The table indicates that $350/800 = 44\%$ of the males are admitted, compared with only $200/600 = 33\%$ of the females. One way to express the disparity is to subtract the two percentages: $44\% - 33\% = 11$ percentage points. Although such

84. For assistance in coping with percentages, see Zeisel, *supra* note 14, at 1–24.

85. Correlation and regression are discussed in the section titled “Correlation and Regression” below.

86. A table of this sort is called a “cross-tab” or a “contingency table.” Table 5 is “two-by-two” because it has two rows and two columns, not counting rows or columns containing totals.

subtraction is commonly seen in jury discrimination cases,⁸⁷ the difference is inevitably small when the two percentages are both close to zero. If the selection rate for males is 5% and that for females is 1%, the difference is only 4 percentage points. Yet, females have only one-fifth the chance of males of being admitted, and that may be of real concern.

For Table 5, the selection ratio (used by the Equal Employment Opportunity Commission in its “80% rule”) is $33/44 = 75\%$, meaning that, on average, women have 75% the chance of admission that men have.⁸⁸ The analogous statistic used in epidemiology is called the relative risk.⁸⁹ Relative risks are usually quoted as decimals; for example, a selection ratio of 75% corresponds to a relative risk of 0.75.

However, the selection ratio has its own problems. In the last example, if the selection rates are 5% and 1%, then the exclusion rates are 95% and 99%. The ratio is $99/95 = 104\%$, meaning that females have, on average, 104% the risk of males of being rejected. The underlying facts are the same, of course, but this formulation sounds much less disturbing.

A statistic known as the odds ratio is more symmetric. If 5% of male applicants are admitted, the odds on a man being admitted are 5 to 95 = 1 to 19; the odds on a woman being admitted are 1:99. The odds ratio is $(1/99)/(1/19) = 19/99$. The odds ratio for rejection instead of acceptance is the same, except that the order is reversed.⁹⁰ Although the odds ratio has desirable mathematical properties, its meaning may be less clear than that of the selection ratio or the simple difference.⁹¹

Data showing disparate impact are generally obtained by aggregating—putting together—data from a variety of sources. Unless the source material is fairly homogeneous, aggregation can distort patterns in the data. We illustrate

87. *E.g.*, *Woodfox v. Cain*, 772 F.3d 358, 376 (5th Cir. 2014) (presenting percentage-point differences that were seen as establishing a prima facie case of discrimination); Sara Sun Beale, *Grand Jury Law and Practice* §§ 3:18–3:19 (2d ed. update Dec. 2021); David H. Kaye, *Statistical Evidence of Discrimination in Jury Selection*, in *Statistical Methods in Discrimination Litigation* 13 (David H. Kaye & Mikel Aickin eds., 1986).

88. A procedure that selects candidates from the least successful group at a rate less than 80% of the rate for the most successful group “will generally be regarded by the Federal enforcement agencies as evidence of adverse impact.” EEOC Uniform Guidelines on Employee Selection Procedures, 29 C.F.R. § 1607.4(D) (1978). The rule is designed to help spot instances of substantially discriminatory practices, and the commission usually asks employers to justify any procedures that produce selection ratios of 80% or less.

89. *See* Gold et al., *supra* note 15.

90. For women, the odds of rejection are 99 to 1; for men, 19 to 1. The ratio of these odds is 99:19. Likewise, the odds ratio for an admitted applicant being a man as opposed to a denied applicant being a man is 99:19.

91. *But see* Joseph B. Kadane, *Odds Ratios as a Measure of Disproportionate Treatment: Application to Jury Venires*, 21 *Law, Probability & Risk* 163, 172 (2022), <https://doi.org/10.1093/lpr/mgad003> (“[o]dds ratios are a natural and easily interpretable way of summarizing data in cases alleging racial disproportion”).

the problem with the hypothetical admission data in Table 5. Applicants can be classified not only by gender and admission but also by the college to which they applied, as in Table 6.

Table 6. Admissions by Sex and College

Decision	Engineering		Business	
	Male	Female	Male	Female
Admit	300	100	50	100
Deny	300	100	150	300

The entries in Table 6 add up to the entries in Table 5. Expressed in a more technical manner, Table 5 is obtained by aggregating the data in Table 6. Yet there is no association between sex and admission in either college; within each college, men and women are admitted at identical rates. Combining two colleges with no association produces a university in which sex is associated strongly with admission. The explanation for this paradox is that the business college, to which most of the women applied, admits relatively few applicants. It is easier to be accepted at the engineering college, the college to which most of the men applied. Combining data from heterogeneous sources (two very different colleges) has made the selectivity of the college into a confounding variable.⁹²

Does a Graph Portray Data Fairly?

Graphs are useful for revealing key characteristics of a batch of numbers, trends over time, and the relationships among variables.⁹³

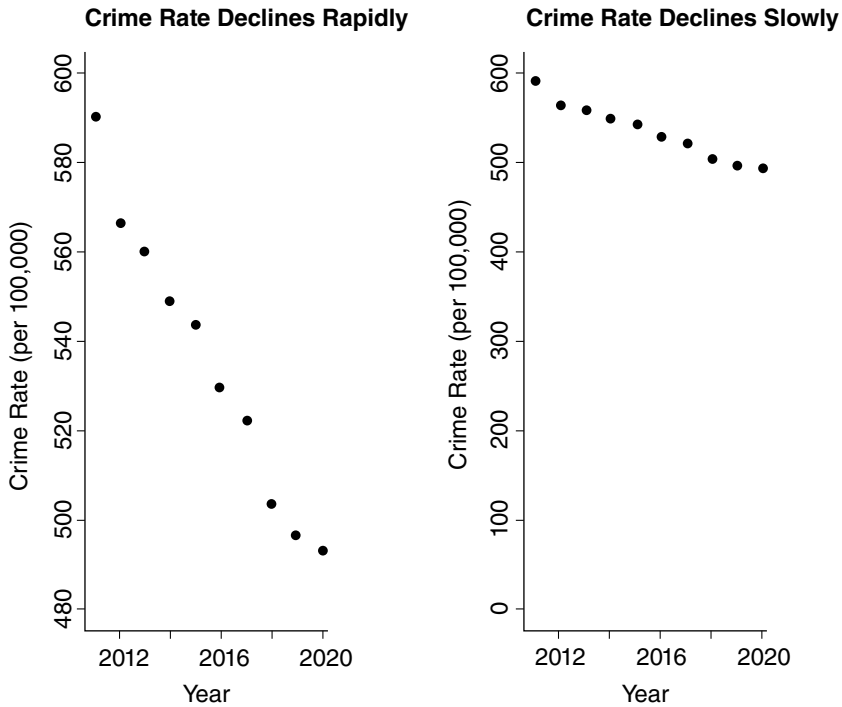
How Are Trends Displayed?

Graphs that plot values over time are useful for seeing trends. However, the scales on the axes matter. In Figures 1 and 2, the rate of all crimes of domestic violence in Florida (per 100,000 people) appears to decline rapidly over the 10 years from

92. Tables 5 and 6 are hypothetical, but closely patterned on a real example. See P.J. Bickel et al., *Sex Bias in Graduate Admissions: Data from Berkeley*, 187 Science 398 (1975), <https://doi.org/10.1126/science.197.4175.398>. The tables are an instance of Simpson’s paradox.

93. See generally Alberto Cairo, *The Truthful Art: Data, Charts, and Maps for Communication* (2016); Alberto Cairo, *The Functional Art: An Introduction to Information Graphics and Visualization* (2012); William S. Cleveland, *The Elements of Graphing Data* (rev. ed. 1994); Edward R. Tufte, *The Visual Display of Quantitative Information* (2d ed. 2001).

Figures 1 and 2. Manipulating the scale of a graph.



2011 through 2020; in Figure 2, the same rate appears to drop slowly.⁹⁴ The moral is simple: Pay attention to the markings on the axes to determine whether the scale is appropriate. A similar admonition applies to bar charts, which display counts or rates across categories.⁹⁵

How Are Distributions Displayed?

A graph commonly used to display the distribution of data is the histogram. One axis denotes the numbers, and the other indicates how often these numbers fall within specified intervals (called “bins” or “class intervals”). For example, we

94. Florida Statistical Analysis Center, Florida Dep’t of Law Enforcement, Statewide Reported Domestic Violence Offenses in Florida, 1992–2020, <https://perma.cc/KQ7F-H6UM>. The data are from the Florida Uniform Crime Report statistics on crimes ranging from simple stalking and forcible fondling to murder and arson.

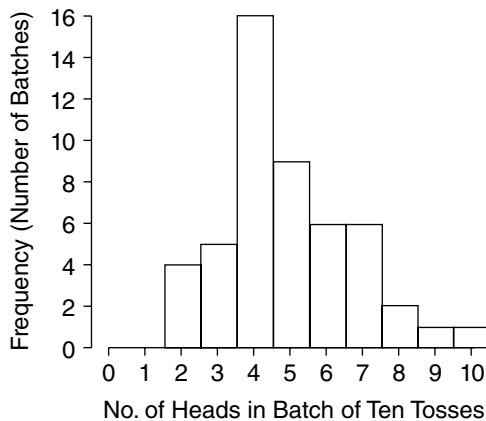
95. See David Spiegelhalter, *The Art of Statistics: Learning from Data* 25–26 (2019) (for an example using survival rates of patients admitted to different hospitals).

flipped a quarter 10 times in a row and counted the number of heads in this “batch” of 10 tosses. Repeating this exercise until we obtained 50 batches, we recorded the following counts:⁹⁶

7 7 5 6 8 4 2 3 6 5 4 3 4 7 4 6 8 4 7 4 7 4 5 4 3
4 4 2 5 3 5 4 2 4 4 5 7 2 3 5 4 6 4 9 10 5 5 6 6 4

The histogram is shown in Figure 3.⁹⁷ A histogram displays how the data are distributed over the range of possible values. The spread can be made to appear larger or smaller, however, by changing the scale of the horizontal axis. Likewise, the shape can be altered somewhat by changing the size of the bins. It may be worth inquiring how the analyst chose the bin widths. As the width of the bins decreases, the graph becomes more detailed, but the appearance becomes

Figure 3. Histogram showing how frequently various numbers of heads appeared in 50 batches of 10 tosses of a quarter.



96. The coin landed heads 7 times in the first 10 tosses; by coincidence, there were also 7 heads in the next 10 tosses; there were 5 heads in the third batch of 10 tosses; and so forth.

97. In Figure 3, the bin width is 1. There were no 0's or 1's in the data, so the bars over 0 and 1 disappear. There is a bin from 1.5 to 2.5; the four 2's in the data fall into this bin, so the bar over the interval from 1.5 to 2.5 has height 4. There is another bin from 2.5 to 3.5, which catches five 3's; the height of the corresponding bar is 5. And so forth. Five is the most likely count in any random batch, but in these 50 batches, a count of 4 heads occurred more frequently, as shown by the larger height for bin 4.

All the bins in Figure 3 have the same width, so this histogram is just like a bar graph. However, data are often published in tables with unequal intervals. The resulting histograms will have unequal bin widths; bar heights should be calculated so that the areas (height $\times$ width) are proportional to the frequencies. In general, a histogram differs from a bar graph in that it represents frequencies by area, not height. See Freedman et al., *supra* note 14, at 31–41.

more ragged until finally the graph is effectively a plot of each datum. The optimal bin width depends on the subject matter and the goal of the analysis.

Is an Appropriate Measure Used for the Center of a Distribution?

Perhaps the most familiar descriptive statistic is the mean (or “arithmetic mean”). The mean can be found by adding all the numbers and dividing the total by how many numbers were added. By comparison, the median cuts the numbers into halves: half the numbers are larger than the median and half are smaller.⁹⁸ Yet a third statistic is the mode, which is the most common number in the dataset. These statistics are different, although they are not always clearly distinguished. In ordinary language, the arithmetic mean, the median, and the mode seem to be referred to interchangeably as “the average.” In statistical parlance, however, the average is the arithmetic mean. The mode is rarely used by statisticians, because it is unstable: Small changes to the data often result in large changes to the mode. The mean takes account of all the data—it involves the total of all the numbers; however, particularly with small datasets, a few unusually large or small observations may have too much influence on the mean. The median is resistant to such outliers.

Studies of damage awards in tort cases find that the mean is larger than the median.⁹⁹ This is because the mean takes into account (indeed, is heavily influenced by) the magnitudes of the relatively few very large awards, whereas the median is influenced only by the number of such awards. If one is seeking a single, representative number for the awards, the median may be more useful than the mean.¹⁰⁰ Still, if the issue is whether insurers were experiencing more

98. Technically, at least half the numbers are at the median or larger; at least half are at the median or smaller. When the distribution is symmetric, the mean equals the median. The values diverge, however, when the distribution is asymmetric, or skewed.

99. Herbert M. Kritzer et al., *An Exploration of “Noneconomic” Damages in Civil Jury Awards*, 55 Wm. & Mary L. Rev. 971 (2014) (summarizing empirical research); Thomas H. Cohen & Steven K. Smith, U.S. Dep’t of Justice, Bureau of Justice Statistics Bulletin NCJ 202803, Civil Trial Cases and Verdicts in Large Counties 2001, 10 (2004) (in a probability sample of cases, the median compensatory award in wrongful death cases was \$961,000, whereas the mean award was around \$3.75 million for the 162 cases in which the plaintiff prevailed); cf. Stephen J. Choi & Theodore Eisenberg, *Punitive Damages in Securities Arbitration: An Empirical Study*, 39 J. Legal Stud. 497, 513 tbl. 1 (2010) (reporting much higher mean than median arbitration awards). In *TXO Production Corp. v. Alliance Resources Corp.*, 509 U.S. 443 (1993), briefs portraying the punitive damage system as out of control pointed to mean punitive awards. These were some 10 times larger than the median awards described in briefs defending the system of punitive damages. Michael Rustad & Thomas Koenig, *The Supreme Court and Junk Social Science: Selective Distortion in Amicus Briefs*, 72 N.C. L. Rev. 91, 145–47 (1993).

100. E.g., *Heinrich v. Sweet*, 308 F.3d 48 (1st Cir. 2002) (two outliers made the median survival time a better indicator of what might have happened had a patient not undergone

costs from jury verdicts, the mean is the more appropriate statistic: The total of the awards is directly related to the mean, not to the median.¹⁰¹

Research also has shown considerable stability in the ratio of punitive to compensatory damage awards, and the Supreme Court has placed great weight on this ratio in deciding whether punitive damages are excessive in a particular case. In *Exxon Shipping Co. v. Baker*,¹⁰² Exxon contended that an award of \$2.5 billion in punitive damages for a catastrophic oil spill in Alaska was unreasonable under federal maritime law. The Court looked to a “comprehensive study of punitive damages awarded by juries in state civil trials [that] found a median ratio of punitive to compensatory awards of just 0.62:1, but a mean ratio of 2.90:1.”¹⁰³ The higher mean could be the result of some large and atypical punitive awards.¹⁰⁴ Looking to the median ratio as “the line near which cases like this one largely should be grouped,” the majority concluded that “a 1:1 ratio, which is above the median award, is a fair upper limit in such maritime cases [of reckless conduct].”¹⁰⁵

Is an Appropriate Measure of Variability Used?

The location of the center of a batch of numbers reveals nothing about the variations exhibited by these numbers. The numbers 1, 2, 5, 8, 9 have 5 as their mean and median. So do the numbers 5, 5, 5, 5, 5. In the first batch, the numbers vary considerably about their mean; in the second, the numbers do not vary at all.

experimental therapy). In passing on proposed settlements in class-action lawsuits, courts have been advised to look to the magnitude of the settlements negotiated by the parties. But the mean settlement will be large if a higher number of meritorious, high-cost cases are resolved early in the life cycle of the litigation. This possibility led the court in *In re Educational Testing Service Praxis Principles of Learning and Teaching, Grades 7–12 Litigation*, 447 F. Supp. 2d 612, 625 (E.D. La. 2006), to regard the smaller median settlement as “more representative of the value of a typical claim than the mean value” and to use this median in extrapolating to the entire class of pending claims.

101. To get the total award, just multiply the mean by the number of awards; by contrast, the total cannot be computed from the median. (The more pertinent figure for the insurance industry is not the total of jury awards, but actual claims experience including settlements; of course, even the risk of large punitive damage awards may have considerable impact.)

102. 554 U.S. 471 (2008).

103. *Id.* at 499.

104. According to the Court, “the outlier cases subject defendants to [disproportionate] punitive damages,” *id.* at 500, and the “stark unpredictability” of these rare awards is the “real problem.” *Id.* at 499. This perceived unpredictability has been the subject of various statistical studies and much debate. See Theodore Eisenberg et al., *Variability in Punitive Damages: Empirically Assessing Exxon Shipping Co. v. Baker*, 166 J. Institutional & Theoretical Econ. 5 (2010) (also criticizing the use of a constant 1:1 ratio across the entire range of damage awards); Anthony J. Sebok, *Punitive Damages: From Myth to Theory*, 92 Iowa L. Rev. 957 (2007).

105. 554 U.S. at 513.

Statistical measures of variability include the range, the interquartile range, and the standard deviation. The range is the difference between the largest number in the batch and the smallest. The range seems natural, and it indicates the maximum spread in the numbers, but the range is unstable because it depends entirely on the most extreme values.

The interquartile range is the difference between the 25th and 75th percentiles. By definition, 25% of the data fall below the 25th percentile, 90% fall below the 90th percentile, and so on. The median is the 50th percentile. The interquartile range contains 50% of the numbers and is resistant to changes in extreme values.

The standard deviation can be viewed as a kind of average or typical deviation from the mean.¹⁰⁶ Suppose we have a batch of 50 numbers whose mean is 100. The standard deviation is found by subtracting this mean from each number, squaring each of these deviations, adding up the squared deviations, dividing by 50 (to get the mean squared deviation, or “variance”), and taking the square root. For example, if one of the numbers is 105, then it deviates from the mean by 5, and the square of 5 is $5^2 = 25$. If the mean of all such squared deviations is 400, then the standard deviation is the square root of 400, which is 20. Taking the square root gets back to the original scale of the measurements. For example, if the numbers are measurements of length in inches, the mean and standard deviation are also in inches.

There are no hard and fast rules about which statistic is the best. In general, the bigger the measures of spread are, the more the numbers are dispersed.¹⁰⁷ Particularly in small datasets, the standard deviation can be influenced heavily by a few outlying values. To assess the extent of this influence, the mean and the standard deviation can be recomputed with the outliers discarded. These “trimmed” statistics (and some others) are more robust (less sensitive to outliers). Beyond this, any of the statistics can (and often should) be supplemented with a diagram that displays much of the data.

106. When the distribution follows the normal curve, about 68% of the data will lie within plus-or-minus 1 standard deviation of the mean; about 95% will lie within 2 standard deviations of the mean. For other distributions, the proportions will be different.

107. In *Exxon Shipping Co. v. Baker*, 554 U.S. 471 (2008), along with the mean and median ratios of punitive to compensatory awards of 0.62 and 2.90, the Court referred to a standard deviation of 13.81. *Id.* at 499. These numbers led the Court to remark that “[e]ven to those of us unsophisticated in statistics, the thrust of these figures is clear: the spread is great, and the outlier cases subject defendants to punitive damages that dwarf the corresponding compensatories.” *Id.* at 500. The size of the standard deviation compared to the mean supports the observation that ratios in the cases of jury award studies are dispersed. A graph of each pair of punitive and compensatory damages offers more insight into how scattered these figures are. See Theodore Eisenberg et al., *The Predictability of Punitive Damages*, 26 J. Legal Stud. 623 (1997), and the section titled “Scatter Diagrams” below.

What Inferences Can Be Drawn from the Data?

The inferences that may be drawn from a study depend on the design of the study and the quality of the data (see section titled “How Have the Data Been Collected?” above). The data might not address the issue of interest, might be systematically in error, or might be difficult to interpret because of confounding. Statisticians would group these concerns together under the rubric of “bias.” In this context, bias means systematic error, with no connotation of prejudice. We turn now to another concern, namely, the impact of random chance on study results. This contribution to total error may be called random error, sampling error, chance error, or statistical error.¹⁰⁸

If a pattern in the data is the result of chance, it is likely to wash out when more data are collected. By applying the laws of probability, a statistician can assess the likelihood that random error will create spurious patterns of certain kinds. Such assessments are often viewed as essential when making inferences from data. Thus, statistical inference typically involves tasks such as the following, which will be discussed in the rest of this guide.

- *Estimation.* A statistician draws a sample from a population (see section titled “Descriptive Surveys and Censuses” above) and estimates a parameter—that is, a numerical characteristic of the population. Random error will throw the estimate off the mark. The question is, by how much? The precision of an estimate is usually reported in terms of the standard error and a confidence interval.
- *Significance testing.* A “null hypothesis” is formulated—for example, that a parameter takes a particular value. Because of random error, an estimated value for the parameter is likely to differ from the value specified by the null—even if the null is right. (“Null hypothesis” is often shortened to “null.”) How likely is it to get a difference as large as, or larger than, the one observed in the data? This chance is known as a *p*-value. Small *p*-values argue against the null hypothesis as such values suggest the observed difference is not likely due to chance alone. Statistical significance can be determined by reference to the *p*-value; significance testing is the technique for computing *p*-values and determining statistical significance. Significance testing is sometimes called hypothesis testing; some statisticians

108. Econometricians use the parallel concept of random disturbance terms. See Rubinfeld & Card, *supra* note 25. Randomness and cognate terms have precise technical meanings; it is randomness in the technical sense that justifies the probability calculations behind standard errors, confidence intervals, and *p*-values (see section titled “What Does It Mean to Be Random?” above and the sections titled “Estimation” and “*p*-values, Significance Levels, and Hypothesis Tests” below). For a discussion of samples and populations, see section titled “Descriptive Surveys and Censuses” above.

restrict the latter term to instances in which there are two explicit hypotheses to be addressed.¹⁰⁹

- *Developing a statistical model.* Statistical inferences often depend on the validity of statistical models for the data. If the data are collected on the basis of a probability sample or a randomized experiment, there will be statistical models that suit the occasion, and inferences based on these models will be secure. Otherwise, calculations are generally based on analogy: This group of people is like a random sample; that observational study is like a randomized experiment. The fit between the statistical model and the data-collection process may then require examination—how good is the analogy? If the model breaks down, that will bias the analysis.
- *Computing posterior probabilities.* Given the sample data, what is the probability that a particular hypothesis about the population (such as the null hypothesis in a significance test) is true? The question might be of direct interest to the courts, especially when translated into English; for example, the null hypothesis might be that African-American and white job applicants have the same probability of passing a qualifying examination—in other words, the test does not adversely impact one group. Posterior probabilities can be computed using a formula called Bayes' rule.¹¹⁰

109. As explained in the section titled “*p*-values, Significance Levels, and Hypothesis Tests” below, tests of significance focus on the degree to which data are inconsistent with a specified null hypothesis. There is no explicit alternative hypothesis, although frequently an alternative is implicit in the choice to use a one-sided or two-sided test. The *p*-value is the probability that data as or more unexpected than the data observed would occur by chance under the null hypothesis. The statistical significance of the result is assessed in terms of the *p*-value. Sir Ronald Fisher usually is credited with introducing this framework. *But see* Michael Cowles & Caroline Davis, *On the Origins of the .05 Level of Statistical Significance*, 37 *Am. Psych.* 553 (1982).

By contrast, formal hypothesis testing (as developed by Jerzy Neyman and Egon Pearson) requires explicit specification of an alternative hypothesis and then development of a test procedure with pre-specified probabilities of making two types of errors—type I (rejecting the null hypothesis when it is true) and type II (failing to reject the null hypothesis when the alternative is true). The end result here is a decision rather than a *p*-value. The two procedures end up being closely related in practice despite their philosophical differences. *See, e.g.,* Erich L. Lehmann, *The Fisher, Neyman-Pearson Theories of Testing Hypotheses: One Theory or Two?*, 88 *J. Am. Stat. Ass'n* 1242 (1993), <https://doi.org/10.2307/2291263>.

110. The theorem is named after the Reverend Thomas Bayes (England, c. 1701–1761). An elementary version of the rule is derived in the Appendix. Bayes' essay on the subject was published after his death: *An Essay Toward Solving a Problem in the Doctrine of Chances*, 53 *Phil. Trans. Royal Soc'y London* 370 (1763–1764). On the foundations and varieties of Bayesian and other forms of statistical inference, *see, for example,* Richard M. Royall, *Statistical Inference: A Likelihood Paradigm* (1997); David Freedman, *Some Issues in the Foundation of Statistics*, 19 *Found. Sci.* 19 (1995), *reprinted in* *Topics in the Foundation of Statistics* 19 (Bas C. van Fraassen ed., 1997); Hal S. Stern, *Comparing Philosophies of Statistical Inference*, in *Handbook of Forensic Statistics* 91 (David Banks et al. eds., 2021); *see also* David H. Kaye, *What Is Bayesianism?* in *Probability and Inference in the Law of Evidence: The Uses and Limits of Bayesianism* (Peter Tillers & Eric Green eds., 1988), *reprinted in*

Key ideas of estimation and testing will be illustrated by courtroom examples, with some complications and mathematical details omitted for ease of presentation.¹¹¹ Bayesian reasoning with regard to forensic-science testimony is discussed in the Appendix.

The first example, on estimation, concerns the Presidential Recordings and Materials Preservation Act of 1974, which impounded President Richard Nixon's presidential papers after he resigned.¹¹² Nixon sued, seeking compensation on the theory that the materials belonged to him personally. Courts ruled in his favor: Nixon was entitled to the fair market value of the papers, with the amount to be proved at trial.¹¹³

The Nixon papers were stored in 20,000 boxes at the National Archives in Alexandria, Virginia. It was plainly impossible to value this entire population of material. Appraisers for the plaintiff therefore took a random sample of 500 boxes. (From this point on, details are simplified; thus, the example becomes somewhat hypothetical.) The appraisers determined the fair market value of each sample box. The average of the 500 sample values turned out to be \$2,000. The standard deviation (see section titled "Is an Appropriate Measure of Variability Used?" above) of the 500 sample values was \$2,200. Many boxes had low appraised values, whereas some boxes were considered to be extremely valuable; this spread explains the large standard deviation.

Estimation

What Estimator Should Be Used?

With the Nixon papers, it is natural to use the average value of the 500 sample boxes to estimate the average value of all 20,000 boxes comprising the population. With the average value for each box having been estimated as \$2,000, the plaintiff demanded compensation in the amount of $20,000 \times \$2,000 = \$40,000,000$.

In more complex problems, statisticians may have to choose among several estimators. Generally, estimators that tend to make smaller (or less costly) errors are preferred; however, "error" (or the losses that result from errors) might be quantified in more than one way. Moreover, the advantage of one estimator over another may depend on features of the population that are largely unknown, at least before the data are collected and analyzed. For complicated problems,

28 *Jurimetrics J.* 161 (1988) (distinguishing between "Bayesian probability," "Bayesian statistical inference," "Bayesian inference writ large," and "Bayesian decision theory").

111. Some technical details appear in the appendices to David H. Kaye & David A. Freedman, *Reference Guide on Statistics*, in *Reference Manual on Scientific Evidence* (3d ed. 2011).

112. 44 U.S.C. § 2111.

113. *Nixon v. United States*, 978 F.2d 1269 (D.C. Cir. 1992); *Griffin v. United States*, 935 F. Supp. 1 (D.D.C. 1995).

professional skill and judgment may therefore be required when choosing a sample design and an estimator. In such cases, the choices and the rationale for them should be documented.

What Is the Standard Error?

An estimate based on a sample is likely to be off the mark, at least by a small amount, because of random error. The standard error gives the likely magnitude of this random error, with smaller standard errors indicating better estimates.¹¹⁴ In our example of the Nixon papers, the standard error for the sample average can be computed from (1) the size of the sample—500 boxes—and (2) the standard deviation of the sample values. Bigger samples give estimates that are more precise. Accordingly, the standard error should go down as the sample size grows, although the rate of improvement slows as the sample gets bigger. (“Sample size” and “the size of the sample” just mean the number of items in the sample; the “sample average” is the average value of the items in the sample.) The standard deviation of the sample comes into play by measuring heterogeneity. The less heterogeneity in the values, the smaller the standard error. For example, if all the values were about the same, a tiny sample would give an accurate estimate. Conversely, if the values are quite different from one another, a larger sample would be needed.

With a random sample of 500 boxes and a standard deviation of \$2,200, the standard error for the sample average is estimated to be about \$100.¹¹⁵ The plaintiff’s total demand was figured as the number of boxes (20,000) times the sample average (\$2,000), or \$40,000,000. Therefore, the standard error for the total

114. We distinguish between (1) the standard deviation of the sample data, which measures the spread in the sample values, and (2) the standard error of the sample average, which measures the likely size of the random error in the sample average. Generally, the standard error of an estimator (such as the sample average) is the standard deviation of that estimator as applied to (hypothetically) repeated samples. Courts typically use the broader term “standard deviation” when referring to the standard error. *E.g.*, *Castaneda v. Partida*, 430 U.S. 482 (1977).

115. The standard error for the sample average equals

$$\sqrt{\frac{N-n}{N-1}} \times \frac{\sigma}{\sqrt{n}}.$$

Freedman et al., *supra* note 14, at 367–70. *N* stands for the size of the population, which is 20,000; *n* stands for the size of the sample, which is 500. The first factor, with the *N*’s in it, is the finite sample correction factor. Here, as in many other such examples, the correction factor is so close to 1 that it can safely be ignored. (This is why the size of population usually has no bearing on the precision of the sample average as an estimator for the population average.) Next, σ is the population standard deviation. Its value is unknown but can be estimated by the sample standard deviation of \$2,200. The standard error for the sample mean is therefore estimated from the data as \$2,200/ $\sqrt{500}$, which is nearly \$100.

demand is 20,000 times the standard error for the sample average: $20,000 \times \$100 = \$2,000,000$.¹¹⁶

How is the standard error to be interpreted? Just by the luck of the draw, a few too many high-value boxes may have come into the sample, in which case the estimate of \$40,000,000 is too high. Or, a few too many low-value boxes may have been drawn, in which case the estimate is too low. This is random error. The net effect of random error is unknown, because data are available only on the sample, not on the full population. However, the net effect is likely to be something close to the standard error of \$2,000,000. Random error throws the estimate off, one way or the other, by something close to the standard error. The role of the standard error is to gauge the likely size of the random error.

The plaintiff's argument may be open to a variety of objections, particularly regarding appraisal methods. However, the sampling plan is sound, as is the extrapolation from the sample to the population. And there is little need for a larger sample: Relative to the total claim of \$40 million, a standard error of \$2 million is reasonably small.

What Is the Confidence Interval?

Although random errors larger in magnitude than the standard error are commonplace, random errors larger in magnitude than two or three times the standard error are unusual. Confidence intervals make these ideas more precise. Usually, a confidence interval for the population average (the average of all the values in the population) is centered at the sample average; the desired confidence level is obtained by adding and subtracting a suitable multiple of the standard error. In dealing with large samples, statisticians who say that the population average falls within 1 standard error of the sample average will be correct about 68% of the time. Those who say "within 2 standard errors" will be correct about 95% of the time, and those who say "within 3 standard errors" will be correct about 99.7% of the time, and so forth.

The normal curve and large samples

These confidence levels correspond to areas under a famous bell-shaped curve—the normal curve. According to a fundamental theorem of statistics (the central

116. We are assuming a simple random sample. Generally, the formula for the standard error must take into account the method used to draw the sample and the nature of the estimator. In fact, the Nixon appraisers used more elaborate statistical procedures. Moreover, they valued the material as of 1995, extrapolated backward to the time of taking (1974), and then added interest. After 20 years of litigation, Nixon's estate settled with the government for \$18 million. U.S. Dep't of Just., Press Release, June 12, 2000, <https://perma.cc/PJC9-47CY>.

limit theorem), if we were to draw not merely a single large sample from a much larger population, or even 500 of them as in the Nixon example, but millions upon millions of them (replacing the items from each sample before drawing the next one), and if we then sorted the sample averages into appropriate bins of a histogram, the heights of the bins would be greatest near the population average, and they would fall off symmetrically on each side as prescribed by the values of the normal curve centered at this point and having a standard deviation that is the standard error.¹¹⁷

The confidence interval relies on this understanding of the distribution of sample means but goes in the other direction. Knowing how the sample statistic behaves as a function of the population parameters, we draw an interval around the sample statistic that we hope will cover the true value (the parameter). The mathematical properties of the normal distribution justify the numbers used to express the “confidence.”¹¹⁸ In particular,

- To get a 68% confidence interval, start at the sample average, then add and subtract 1 standard error.
- To get a 95% confidence interval, start at the sample average, then add and subtract twice the standard error.
- To get a 99.7% confidence interval, start at the sample average, then add and subtract three times the standard error.

With the Nixon papers, the 68% confidence interval for plaintiff’s total demand runs

- from $\$40,000,000 - \$2,000,000 = \$38,000,000$
- to $\$40,000,000 + \$2,000,000 = \$42,000,000$.

117. The normal curve is the density of a normal distribution. The distribution has two parameters—the population mean (often denoted μ) and the standard deviation (σ), which is the standard error of the variable x (here, the sample mean, which varies from one sample to the next). The equation is

$$f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

where $e = 2.71828 \dots$ and $\pi = 3.14159 \dots$. Once the two parameters μ and σ are given, the density is completely specified.

118. The area under the normal curve between $x = \mu - \sigma$ and $x = \mu + \sigma$ is close to 68.3%. Likewise, the area between $\mu - 2\sigma$ and $\mu + 2\sigma$ is close to 95.4%. Many academic statisticians would use ± 1.96 SE for a 95% confidence interval. However, the normal curve only gives an approximation to the relevant chances, and the error in that approximation will often be larger than a few tenths of a percent. For simplicity, we use ± 1 SE for the 68% confidence level, and ± 2 SE for 95% confidence.

The 95% confidence interval runs

- from $\$40,000,000 - (2 \times \$2,000,000) = \$36,000,000$
- to $\$40,000,000 + (2 \times \$2,000,000) = \$44,000,000$.

The 99.7% confidence interval runs

- from $\$40,000,000 - (3 \times \$2,000,000) = \$34,000,000$
- to $\$40,000,000 + (3 \times \$2,000,000) = \$46,000,000$.

To write this more compactly, we abbreviate standard error as SE. Thus, 1 SE is one standard error, 2 SE is twice the standard error, and so forth. With a large sample and an estimate like the sample average, a 68% confidence interval ranges from

estimate – 1 SE to estimate + 1 SE.

A 95% confidence interval is ranges from

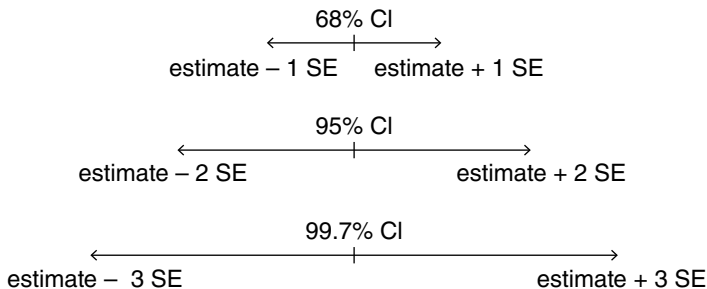
estimate – 2 SE to estimate + 2 SE.

A 99.7% confidence interval ranges from

estimate – 3 SE to estimate + 3 SE.

For a given sample size, increased confidence can be attained only by widening the interval. The 95% confidence level is the most popular, but some authors use 99%, and 90% is seen on occasion. (The corresponding multipliers on the SE are about 2, 2.6, and 1.6, respectively.) The phrase “margin of error” generally means twice the standard error. In medical journals, “confidence interval” is often abbreviated as “CI.” The relationship between width and confidence is shown in Figure 4.

Figure 4. Confidence coefficients for CIs of ± 1 , 2, and 3 standard errors of a normally distributed estimator.



The picture shows a tradeoff between precision (the width of the interval) and confidence (that the interval covers the actual population value). In sum, an estimate based on a sample will differ from the exact population value, because of random error. The standard error gives the likely size of the random error. If the standard error is small, random error probably has little effect. If the standard error is large, the estimate may be far away from the population value. Confidence intervals are a technical refinement that provides a formalism for assessing the impact of random error or chance variation on the estimate.

Other situations

Intervals based on the standard error, with confidence levels read off the normal curve, are appropriate for estimators that are essentially unbiased and obey the central limit theorem. They generally work for sums, averages, and rates, although much depends on the design of the sampling and other matters. CIs determined via the normal curve may not work well as estimates of extremely small quantities. Table 1 of the section titled “Is the Measurement Process Valid?” presents an extreme example. The table showed that toolmark examiners given 30 pairs of cartridges from ammunition fired from the same guns and 45 pairs from different guns correctly classified every pair according to the true source. But what if the same examiners had been brought back for a second experiment with the same cartridge cases (and with no memory of the first experiment)? Would they have done as well? What about a third experiment? It would be extraordinary if they never erred in one experiment after another *ad infinitum*. Perhaps we can think of the one experiment as if it were a random sample from an infinite population of repeated experiments. Given this framework, we might want a confidence interval that will estimate the proportion of false-positive errors on the part of the examiners in an infinite stream of experiments with guns and ammunition exactly like those in the study. The observed proportion of errors in the one experiment is zero, which makes the plug-in estimate of the standard error zero as well. However, a confidence interval of zero width cannot be right. The probability of false-positive errors could be higher than zero, and an examiner could still get all 45 determinations for the false pairs correct. The problem is that because every determination in the sample (the one experiment) is correct, there is no variability to inform what we might see in future samples (experiments).

How should we proceed? It is easy to show that an interval from 0 to 6.5% would achieve at least 95% “confidence.”¹¹⁹ So if we are willing to embrace anything in the range of 0 to 6.5% as the true error-rate when given samples with an

119. When the long-run false-positive error rate is 6.5%, the probability of 45 independent correct classifications of cartridges from different pairs of guns is $(1-0.065)^{45} = 0.049$. When the “true” error rate is greater than 6.5%, that probability for zero errors in the sample is smaller still.

observed rate of 0/45, we will be correct at least 95% of the time (in the long run). But this interval is a very conservative estimate. How to construct a confidence or other interval for the probability of an event when 0/ n are observed goes by the name of the zero-numerator problem, and several approaches have been proposed.¹²⁰

In still other situations (for example, where observations are dependent), other methods can be used to obtain confidence intervals. One class of methods repeatedly resamples from the observed data to approximate what would happen in repeated samples from the full population.¹²¹ With modern computers, the resulting bootstrap confidence intervals are enticingly easy to calculate, but they too require certain assumptions to be verified.¹²²

How Big Should the Sample Be?

There is no easy answer to this sensible question. Much depends on the level of error that is tolerable, the material being sampled, and the sampling method. Increasing the size of the sample provides no protection against bias (“nonsampling error”). Indeed, beyond some point, large samples are harder to manage and more vulnerable to nonsampling error. To reduce bias, the researcher must improve the design of the study or use a statistical model more tightly linked to the data-collection process. Larger samples generally will reduce the level of random error (“sampling error”). It rarely will be sensible to draw a probability sample with fewer than, say, two or three dozen items, and with such small samples, methods based solely on the normal curve (see section titled “What Is the Standard Error?” above) will not apply. A pilot sample is often valuable, partly to obtain an initial estimate of characteristics of the population that may be used in

120. E.g., Noriah M. Al-Kandari & Paul H. Garthwaite, *Bayesian Analysis of Misclassified Binomial Data: Double-Sampling and the Zero-Numerator Problem*, Communications in Statistics—Simulation & Computation (2020), <https://doi.org/10.1080/03610918.2020.1855448>; B. D. Jovanovic & P. S. Levy, *A Look at the Rule of Three*, 51 Am. Statistician 137 (1997); Robert L. Winikler et al., *The Role of Informative Priors in Zero-Numerator Problems: Being Conservative Versus Being Candid*, 56 Am. Statistician 1 (2002).

121. See, e.g., Freedman, *supra* 25, at 147–50; David H. Kaye, *Frequentist Statistical Inference*, in Handbook of Forensic Statistics 39, 66–68 (David Banks et al. eds., 2021) (discussing the basic idea). On the patentability of procedures relying on bootstrapping or other resampling methods, see *SAP America v. InvestPic*, 898 F.3d 1161 (Fed. Cir. 2018) (investment analysis); *Cybergenetics Corp. v. Inst. of Env’t Sci. & Rsch.*, 490 F. Supp. 3d 1237 (N.D. Ohio 2020) (in discerning separate contributors to DNA mixtures), *aff’d*, 856 Fed. App’x 312 (Fed. Cir. 2021).

122. See *In re Sonic Corp. Customer Data Sec. Breach Litig.*, No. 1:17-md-2807, MDL No. 2807, 2021 WL 5916743 (N.D. Ohio Dec. 15, 2021) (bootstrap estimate that losses to credit-card companies from a computer security breach exceeded \$76,000,000 were inadmissible because the bootstrap sample was small and unrepresentative); Bradley Efron & Robert J. Tibshirani, *An Introduction to the Bootstrap* (1994); Freedman, *supra* note 25, at 150.

determining the final sample size.¹²³ If a population appears to be heterogeneous, then alternatives to simple random sampling, such as stratified random samples (that are each fairly homogeneous), may provide more representative samples and more precise estimates for the same total sample size. As these examples illustrate, probability samples require some effort in the design phase.

Population size (i.e., the number of items in the population) usually has little bearing on the precision of estimates for the population average. This is surprising. On the other hand, population size has a direct bearing on estimated totals. Both points are illustrated by the Nixon papers (see section titled “What Is the Standard Error?” above). To be sure, drawing a probability sample from a large population may involve a lot of work. Samples presented in the courtroom have ranged from 5 (tiny) to 1.7 million (huge).¹²⁴

What Are the Technical and Interpretive Difficulties with Confidence Intervals?

To begin with, “confidence” has an esoteric meaning in this context. The confidence level indicates the percentage of the time that intervals from repeated samples would cover the true value. The confidence level does not express the chance that repeated estimates would fall into the stated confidence interval.¹²⁵

123. Suppose a researcher is interested in the association between alcohol intoxication and single-vehicle accidents. A small sample of accident records (along with intuitions) could suggest some guess for the population proportion of single-vehicle crashes in which the driver was found to be intoxicated. Assume the guess is 50%. Suppose further that the researcher is willing to tolerate errors of up to $\pm 3\%$ in the estimate from the larger sample being planned. Finally, suppose that the researcher plans to report a conventional 95% confidence interval. It can be shown that a simple random sample of approximately 1,067 single-vehicle accident records should suffice. Hence, the researcher collects 1,067 records at random and determines the proportion in which intoxication was in the accident record. The observed proportion in this sample might be 40% or 0.4, to pick a concrete number. The earlier guess of 50% no longer matters—it was just used to get a sense of the sample size for the full study, and the researcher now can report a 95% CI from the full study. The estimate becomes $.40 \pm 1.96\sqrt{(.40)(.60)/1067} = .40 \pm .03$, which is about 37% to 43%.

124. *Lebrilla v. Farmers Grp., Inc.*, No. 00-CC-017185 (Cal. Super. Ct., Orange Cnty., Dec. 5, 2006) (preliminary approval of settlement). This was a class action lawsuit on behalf of plaintiffs who were insured by Farmers and had automobile accidents. Plaintiffs alleged that replacement parts recommended by Farmers did not meet specifications: Small samples were used to evaluate these allegations. At the other extreme, it was proposed to adjust Census 2000 for undercount and overcount by reviewing a sample of 1.7 million persons. See Brown et al., *supra* note 36, at 353.

125. Opinions reflecting this misinterpretation include *Turpin v. Merrell Dow Pharm., Inc.*, 959 F.2d 1349, 1353 (6th Cir. 1992) (“If a confidence interval of ‘95 percent between 0.8 and 3.10’ is cited, this means that random repetition of the study should produce, 95 percent of the time, a relative risk somewhere between 0.8 and 3.10.”); *Garcia v. Tyson Foods, Inc.*, 890 F. Supp. 2d 1273, 1285 (D. Kan. 2012) (“Dr. Radwin testified that his study was conducted within a confidence interval of 95—that is ‘if I did this study over and over again, 95 out of a hundred times I would

With the Nixon papers, the 95% confidence interval should not be interpreted as saying that 95% of all random samples will produce estimates in the range from \$36 million to \$44 million. Rather, a high degree of “confidence” makes it reasonable to accept the one sample interval as a plausible statement of what the value of all the papers could be.

Second, the confidence level does not give the probability that the unknown parameter lies within the confidence interval.¹²⁶ For example, the 95% confidence level should not be translated to a 95% probability that the total value of the papers is in the range from \$36 million to \$44 million. According to the frequentist theory of statistics, probability statements cannot be made about population characteristics: Probability statements apply to the behavior of samples. That is why the different term “confidence” is used.

Third, this “confidence” is a statement about the entire region. One might think that values near the center of the interval are more likely to be closer to the true value than those at the extremes. Yet the statement of confidence applies equally to all the values in the interval. Moreover, that the intervals have sharp boundaries does not imply that there is an important difference between a value at the edge of the interval and one just beyond it.

expect to get an average between that interval.”); *In re Silicone Gel Breast Implants Prods. Liab. Litig.*, 318 F. Supp. 2d 879, 897 (C.D. Cal. 2004) (“a margin of error between 0.5 and 8.0 at the 95% confidence level . . . means that 95 times out of 100 a study of that type would yield a relative risk value somewhere between 0.5 and 8.0”).

Language from another reference guide in the previous edition of this *Reference Manual* that is often quoted may inadvertently convey the incorrect impression that a confidence coefficient such as 95% refers to the percentage of results in (hypothetically) repeated studies that would be expected to lie within the interval reported in the study before the court. *See, e.g., Rhyne v. U.S. Steel Corp.*, 474 F. Supp. 3d 733, 744 (W.D.N.C. 2020) (“If a 95% confidence interval is specified, the range encompasses the results we would expect 95% of the time if samples for new studies were repeatedly drawn from the population.” Reference Guide on Epidemiology, at 580.”). However, simulations suggest that the confidence coefficient for a sample mean from a normal population tends to overstate the probability that the next sample mean will lie within the interval. Geoff Cumming & Robert Maillardet, *Confidence Intervals and Replication: Where Will the Next Mean Fall?*, 11 Psych. Methods 217 (2006), <https://doi.org/10.1037/1082-989X.11.3.217> (finding that “[o]n average, a 95% CI will include just 83.4% of future replication means”). The more technically correct statement in the *Silicone Gel* case would be that “the confidence interval of 0.5 to 8.0 means that the relative risk in the population plausibly could fall within this wide range—and that in roughly 95 times out of 100 random samples from the same population, the confidence intervals (however wide they might be) would include the population value (whatever it is).”

126. *See, e.g., Freedman et al., supra* note 14, at 383–86. Consequently, it is misleading to suggest that “[t]he confidence interval means . . . it is 95 percent likely that the actual [value of the parameter being estimated] is somewhere in the interval, somewhere between the low end of the interval and its high end,” *Center for Biological Diversity v. U.S. Fish & Wildlife Serv.*, 342 F. Supp. 3d 968, 977 (N.D. Cal. 2018), or that “a 95-percent median confidence interval of 89.43 to 90.28 percent [means] that the true median is 95-percent likely to fall within that range,” *Cnty. of Douglas v. Nebraska Tax Equalization & Rev. Comm’n*, 894 N.W.2d 308, 316 (Neb. 2017).

Fourth, for a given confidence level, a narrower interval indicates a more precise estimate, whereas a broader interval indicates less precision.¹²⁷ A high confidence level with a broad interval means very little, but a high confidence level for a small interval is impressive, indicating that the random error in the sample estimate is low. For example, take a 95% confidence interval for a damage claim. An interval that runs from \$36 million to \$44 million is more precise than a much wider interval that goes from \$10 million to \$70 million. Statements about confidence without mention of an interval are practically meaningless.¹²⁸

The final point to make is that standard errors and confidence intervals are often derived from statistical models for the process that generated the data. The model usually has parameters—numerical constants describing the population from which samples were drawn. When the values of the parameters are not known, the statistician must work backwards, using the sample data to make estimates. That was the case for valuing the Nixon papers. One parameter is the average value of all 20,000 boxes, and another parameter is the standard deviation of the 20,000 values.¹²⁹ Generally, the probabilities used in drawing inferences are computed from a model with estimated parameter values.

127. See section titled “What Is the Standard Error?” above. In *Cimino v. Raymark Industries, Inc.*, 751 F. Supp. 649 (E.D. Tex. 1990), *rev’d*, 151 F.3d 297 (5th Cir. 1998), the district court drew certain random samples from more than 6,000 pending asbestos cases, tried these cases, and used the results to estimate the total award to be given to all plaintiffs in the pending cases. The court then held a hearing to determine whether the samples were large enough to provide accurate estimates. The court’s expert, an educational psychologist, testified that the estimates were accurate because the samples matched the population on such characteristics as race and the percentage of plaintiffs still alive. *Id.* at 664. However, the matches occurred only in the sense that population characteristics fell within 99% confidence intervals computed from the samples. The court thought that matches within the 99% confidence intervals proved more than matches within 95% intervals. *Id.* This is backward. To be correct in a few instances with a 99% confidence interval is not very impressive—by definition, such intervals are broad enough to ensure coverage approximately 99% of the time.

128. In *Hilao v. Estate of Marcos*, 103 F.3d 767 (9th Cir. 1996), “an expert on statistics . . . testified that . . . a random sample of 137 claims would achieve ‘a 95% statistical probability that the same percentage determined to be valid among the examined claims would be applicable to the totality of [9,541 facially valid] claims filed.’” *Id.* at 782. There is no 95% “statistical probability” that a percentage computed from a sample will be “applicable” to a population. One can compute a confidence interval from a random sample and be 95% confident that the interval covers some parameter. The computation can be done for a sample of virtually any size, with larger samples giving smaller intervals. What is missing from the opinion is a discussion of the widths of the relevant intervals. For the same reason, it is meaningless to testify, as an expert did in *Ayyad v. Sprint Spectrum, L.P.*, No. RG03–121510 (Cal. Super. Ct., Alameda Cnty.) (transcript, May 28, 2008, at 730), that a simple regression equation is trustworthy because the coefficient of the explanatory variable has “an extremely high indication of reliability to more than 99% confidence level.”

129. These parameters can be used to approximate the distribution of the sample average. See section titled “What Is the Standard Error?” above. Regression models and their parameters are discussed in the section titled “Correlation and Regression” below and in Rubinfeld & Card, *supra* note 25.

If the data come from a probability sample or a randomized controlled experiment (see sections titled “Descriptive Surveys and Censuses” and “Is the Study Designed to Investigate Causation?” above), then the statistical model may be connected tightly to the actual data-collection process. In other situations, using the model may be tantamount to assuming that a sample of convenience is like a random sample, or that an observational study is like a randomized experiment. With the Nixon papers, the appraisers drew a random sample, and that justified the statistical calculations—if not the appraised values themselves. In many contexts, the choice of an appropriate statistical model is less than obvious. When a model does not fit the data-collection process, estimates and standard errors will not be probative.

Standard errors and confidence intervals are designed to take account of random errors but not systematic ones such as selection bias or nonresponse bias (see sections titled “What Method Is Used to Select the Units?” and “Of the Units Selected, Which Provide Measurements?” above). For example, after reviewing studies to see whether a particular drug caused birth defects, a court observed that mothers of children with birth defects may be more likely to remember taking a drug during pregnancy than mothers with normal children.¹³⁰ This selective recall would bias comparisons between samples from the two groups of women. Neither the standard error nor the confidence interval for the difference in drug usage between the groups accounts for this bias.¹³¹

p-values, Significance Levels, and Hypothesis Tests

What Is the p-value?

In 1969, Dr. Benjamin Spock came to trial in the U.S. District Court for Massachusetts. The charge was conspiracy to violate the Military Service Act. The jury was drawn from a panel of 350 persons selected by the clerk of the court. The panel included only 102 women—substantially less than 50%—although a majority of the eligible jurors in the community were female. The shortfall in women was especially poignant in this case: “Of all defendants, Dr. Spock, who had given

130. *Brock v. Merrell Dow Pharm., Inc.*, 874 F.2d 307, 311–12 (5th Cir.), *modified*, 884 F.2d 166 (5th Cir. 1989).

131. In *Brock*, the court stated that the confidence interval took account of bias (in the form of selective recall) as well as random error. 874 F.2d at 311–12. This is wrong. Even if the sampling error were nonexistent—which would be the case if one could interview every woman who had a child during the period that the drug was available—selective recall would produce a difference in the percentages of reported drug exposure between mothers of children with birth defects and those with normal children. In this hypothetical situation, the standard error would vanish. Therefore, the standard error could disclose nothing about the impact of selective recall.

wise and welcome advice on child-rearing to millions of mothers, would have liked women on his jury.”¹³²

Can the shortfall in women be explained by the mere play of random chance? To approach the problem, a statistician could formulate and test a null hypothesis. Here, the null hypothesis says that the panel is like 350 persons drawn at random from a large population that is 50% female. The expected number of women drawn would then be 50% of 350, which is 175. The observed number of women is 102. The shortfall is $175 - 102 = 73$. How likely is it to find a disparity this large or larger, between observed and expected values? The probability is called the *p*-value, or *p*.

The *p*-value is the probability of getting data as extreme as, or more extreme than, the actual data—given that the null hypothesis is true. In the example, *p* can be computed from a simple probability model, expressed as a function that gives the probability of every possible outcome for the number of women in a sample. A reasonable model here, known as the binomial distribution, depends on just two parameters: the size *n* of the sample and a constant probability θ of selecting a woman on each draw.¹³³ Using the binomial formula, the probability of a sample with a shortfall or an excess of at least 73 women turns out to be essentially zero.¹³⁴ The discrepancy between the observed and the expected is far too large to explain by random chance. Indeed, even if the panel had included 155 women, the *p*-value would only be around 0.04, or 4%.¹³⁵ (If the population is more than 50% female, *p* will be even smaller.) In short, the jury panel was nothing like a random sample from the community.

Large *p*-values indicate that a disparity can easily be explained by the play of chance: The data fall within the range likely to be produced by chance

132. Hans Zeisel, *Dr. Spock and the Case of the Vanishing Women Jurors*, 37 U. Chi. L. Rev. 1 (1969). Zeisel’s reasoning was different from that presented in this text. The conviction was reversed on appeal without reaching the issue of jury selection. *United States v. Spock*, 416 F.2d 165 (1st Cir. 1965).

133. In mathematical notation, the probability for the number *x* of women in a sample of size *n* is the function $f(x; n, \theta) = n!/[x!(n-x)!] \times \theta^x(1-\theta)^{n-x}$. This binomial formula for *f* is discussed in many texts, including Freedman et al., *supra* note 14, at 255–61. This model is a good choice when the population of potential jurors is so large that the 50–50 split ($\theta = 0.5$) of men and women does not change appreciably as each juror is selected. A different distribution (called a hypergeometric distribution) would be used if that complication were more important. See, e.g., David H. Kaye, *Ruminations on Jurimetrics: Hypergeometric Confusion in the Fourth Circuit*, 26 Jurimetrics J. 215 (1986).

134. The binomial probability of getting exactly 102 women in the sample is $f(x = 102; n = 350, \theta = 1/2)$, which works out to $1/10^{15}$. With the binomial probability function, we can compute the chance of getting 101 women, or 100, or any other particular number *x*. The chance of getting 102 women or fewer is then computed by addition. The chance is about $2/10^{15}$. The number 10^{15} is 1 followed by 15 zeros, that is, a quadrillion. The chance of an equally extreme outcome in the other direction ($x = 248$ or more) is the same. The *p*-value of $4/10^{15}$ is very bad news for the null hypothesis.

135. See Kaye & Freedman, *supra* note 111, at Appendix B.

variation. On the other hand, if p is very small, something other than chance must be involved: The data are far away from the values expected under the null hypothesis. Significance testing often seems to involve multiple negatives. This is because a statistical test is an argument by contradiction. With the Dr. Spock example, the null hypothesis asserts that the jury panel is like a random sample from a population that is 50% female. The data contradict this null hypothesis because the disparity between what is observed and what is expected (according to the null) is too large to be explained as the product of random chance. In a typical jury discrimination case, small p -values help a defendant appealing a conviction by showing that the jury panel is not like a random sample from the relevant population; large p -values hurt the defendant's case. Likewise, in the usual employment context, small p -values help plaintiffs who complain of discrimination—for example, by showing that a disparity in promotion rates is too large to be explained by chance; conversely, large p -values would be consistent with the defense argument that the disparity is just due to chance.

Because p is calculated by assuming that the null hypothesis is correct, p does not give the chance that the null is true. The p -value merely gives the chance of getting evidence against the null hypothesis as strong as or stronger than the evidence at hand. Chance affects the data, not the hypothesis. According to the frequency theory of statistics, there is no meaningful way to assign a numerical probability to the null hypothesis. The correct interpretation of the p -value can therefore be summarized in two lines:

p is the probability of extreme data given the null hypothesis.
 p is not the probability of the null hypothesis given extreme data.¹³⁶

136. Some opinions present a contrary view. *E.g.*, *Berghuis v. Smith*, 559 U.S. 314, 324 n.1 (2010) (noting that statistical analysis “seeks to determine the probability that the disparity between a group’s jury-eligible population and the group’s percentage in the qualified jury pool is attributable to random chance”); *Vasquez v. Hillery*, 474 U.S. 254, 259 n.3 (1986) (“the District Court . . . ultimately accepted . . . a probability of 2 in 1000 that the phenomenon was attributable to chance”); *United States v. Carter*, 750 F.3d 462, 468 n.15 (4th Cir. 2014) (“a p -value of 0.068 indicates that there was only a 6.8% chance that the correlation was due to chance”). Such statements confuse the probability of the kind of outcome observed, which is computed under some model of chance, with the probability that chance is the explanation for the outcome—the “transposition fallacy.”

Instances of the transposition fallacy in criminal cases are collected in Kaye et al., *supra* note 8, §§ 12.8.2(b) & 14.1.2. In *McDaniel v. Brown*, 558 U.S. 120 (2010), for example, a DNA analyst suggested that a random-match probability of 1/3,000,000 implied a .000033 probability that the defendant was not the source of the DNA found on the victim’s clothing. See David H. Kaye, *The Interpretation of DNA Evidence: A Case Study in Probabilities*, in *Science Policy Decision-Making Educational Modules* (Nat’l Acad. Sci., Eng’g & Med. Comm. on Preparing the Next Generation of Policy Makers for Science-Based Decisions ed., 2016), <https://www.nationalacademies.org/our-work/science-policy-decision-making-educational-modules>.

To recapitulate the logic of p -values: If p is small, the observed data are far from what is expected under the null hypothesis—too far to be readily explained by the operations of chance. That discredits the null hypothesis.

Computing p -values requires statistical expertise. Many methods are available, but only some will fit the occasion. Sometimes standard errors will be part of the analysis, other times they will not be.¹³⁷ Sometimes a difference of two standard errors will imply a p -value of about 5%; other times it will not. In general, the p -value depends on the statistical model, the size of the sample, and the sample statistics.

Is a Difference Statistically Significant?

If an observed difference is in the middle of the distribution that would be expected under the null hypothesis, there is no surprise. The sample data are of the type that often would be seen when the null hypothesis is true. The observed difference (or similar statistic) does not fall into a region that is appropriate for rejecting the null hypothesis. It is not “significant,” as statisticians are wont to say. On the other hand, if the sample difference is far from the expected value—according to the null hypothesis—then the sample is unusual. The difference is significant, and the null hypothesis is rejected.

Statistical significance can be determined by comparing p to a preset value, called the significance level.¹³⁸ In this approach, the null hypothesis is rejected when p falls below this level. In practice, statistical analysts typically use levels of

137. The binomial analysis in the *Spock* case did not use standard errors. The standard error there is the square root of $n\theta(1-\theta) = 9.35$. The observed value of 102 is nearly 8 standard errors below the expected value of 175, which is a lot of standard errors. However, we did not have to perform this “standard deviation analysis” (*Berghuis v. Smith*, 559 U.S. 314, 324 n.1 (2010); *infra* note 139) to see that the observed disparity was incompatible with the null hypothesis. The p -value told us that directly. Using a particular number of standard errors to mark statistical significance comes from the tradition of using the normal curve to approximate the distribution of sample statistics such as the sample proportion or mean. See section titled “What Is the Standard Error?” above.

138. It is not necessary to compute the p -value to perform the test for significance. Instead, taking the Neyman–Pearson approach mentioned in note 105, before analyzing the data, one can define a rejection region that keeps the risk of falsely rejecting the null hypothesis at or below a prespecified level (and that maximizes the power of the test to detect a difference if one is present). Statisticians use the Greek letter alpha (α) to denote this level; α gives the chance of getting a result that produces a rejection, assuming that the null hypothesis is true. Thus, α represents the chance of a false rejection of the null hypothesis (also called a false positive, a false alarm, or a Type I error). For example, suppose $\alpha = 5\%$. If investigators do many studies, and the null hypothesis happens to be true in each case, then about 5% of the time they would obtain significant results—and falsely reject the null hypothesis. Inasmuch as the data in the rejection region are all such that $p \leq \alpha$, it is common to describe the test, as we have in the text, in terms of the p -value and to call $\alpha = 0.05$ the significance level.

5% and 1%.¹³⁹ The 5% level is the most common in social science, and an analyst who speaks of significant results without specifying the threshold probably is using this figure. An unexplained reference to highly significant results probably means that p is less than 1%. These levels of 5% and 1% have become icons of science and the legal process. In truth, however, such levels are simply common conventions.¹⁴⁰

Because the term “significant” is merely a label for a certain kind of p -value, significance is subject to the same limitations as the underlying p -value. Thus, significant differences may be evidence that something besides random error is at work. They are not evidence that this something is legally or practically important. Statisticians distinguish between statistical and practical significance to make the point. When practical significance is lacking—when the size of a disparity is negligible¹⁴¹—statistical significance may be of no legal significance.¹⁴²

139. The Supreme Court implicitly referred to this practice in *Castaneda v. Partida*, 430 U.S. 482, 496 n.17 (1977), and *Hazelwood School District v. United States*, 433 U.S. 299, 311 n.17 (1977). In these footnotes, the Court described the null hypothesis as “suspect to a social scientist” when a statistic from “large samples” falls more than “two or three standard deviations” from its expected value under the null hypothesis. Although the Court did not say so, these differences produce p -values of about 5% and 0.3% when the statistic is normally distributed. The Court’s standard deviation is our standard error.

140. Some opinions quote statisticians who urge giving readers p -values instead of making pronouncements of “significant” or “not significant” or who “caution against using statistical significance rigidly” as if these authorities are recommending admission of any relevant and properly executed study, regardless of the p -value. *E.g.*, *In re Urethane Antitrust Litig.*, 166 F. Supp. 3d 501, 508–09 (D.N.J. 2016) (relying on such statements to justify admission of findings with p -values of 7.2%, 19.2%, and, in dictum, 50%). However, declarations that p -values are more informative than pronouncements of “significance” do not mean that unimpressive results (those with large p -values) are admissible as proof of the alternative hypothesis. Large p -values indicate that the observed results are not reliable evidence for the alternative hypothesis; consequently, under Rules 403 and 702, very large p -values go to the admissibility as well as the weight of the statistical evidence.

141. Context affects judgments of what outcomes lack practical significance. Some relatively small differences can have a large practical effect. For example, a loss of half a percentage point in the revenue of a corporation that is a consequence of a competitor’s illegal conduct would be practically significant in computing damages if the total revenue was large.

142. *E.g.*, *Brnovich v. Democratic Nat’l Comm.*, 141 S. Ct. 2321, 2343 n.17 (2021) (quoting substantially identical text in the third edition of this Manual); *id.* at 2358 n.4 (dissenting opinion agreeing that “there may be some threshold of what is sometimes called ‘practical significance’—a level of inequality that, even if statistically meaningful, is just too trivial for the legal system to care about”) (dissenting opinion); *Waisome v. Port Auth.*, 948 F.2d 1370, 1376 (2d Cir. 1991) (“though the disparity was found to be statistically significant, it was of limited magnitude”); *United States v. Hernandez-Estrada*, 749 F.3d 1154, 1165 (9th Cir. 2014) (en banc) (“[T]he challenging party must establish not only statistical significance, but also legal significance. . . . In other words, if a statistical analysis shows underrepresentation, but the underrepresentation does not substantially affect the representation of the group in the actual jury pool, then the underrepresentation does not have legal significance in the fair cross-section context.”); *Apsley v. Boeing Co.*, 691 F.3d 1184 (10th Cir. 2012) (a reported p -value of 1/50,000 with respect to hiring older workers was not sufficient to defeat defendant’s motion for summary judgment when the difference between the expected and

It is easy to mistake the p -value for the probability of the null hypothesis given the data (see section titled “What Is the p -value?” above). Likewise, if results are significant at the 5% level, it is tempting to conclude that the null hypothesis has only a 5% chance of being correct.¹⁴³ This temptation should be resisted. From the frequentist perspective, statistical hypotheses are either true or false. Probabilities govern the samples, not the models and hypotheses. The significance level tells us what is likely to happen when the null hypothesis is correct; it does not tell us the probability that the hypothesis is true. Significance comes no closer to expressing the probability that the null hypothesis is true than does the underlying p -value.

Recent Emphasis on the Limitations of p -values

For many of the reasons mentioned above, there has been considerable controversy about issues related to p -values. In response to the perception that many published studies reporting significance with the .05 criterion are not reproducible—the much bruited “replication crisis”¹⁴⁴ in psychology, medicine, and other

actual number hired was only about 50 out of nearly 7,300); *United States v. Henderson*, 409 F.3d 1293, 1306 (11th Cir. 2005) (regardless of statistical significance, excluding law enforcement officers from jury service does not have a large enough impact on the composition of grand juries to violate the Jury Selection and Service Act); *cf. Thornburg v. Gingles*, 478 U.S. 30, 53–54 (1986) (repeating the district court’s explanation of why “the correlation between the race of the voter and the voter’s choice of certain candidates was [not only] statistically significant,” but also “so marked as to be substantively significant, in the sense that the results of the individual election would have been different depending upon whether it had been held among only the white voters or only the black voters”); cases cited, *infra* notes 149–150. *But see Jones v. City of Boston*, 752 F.3d 38, 53 (1st Cir. 2014) (in Title VII cases, “a plaintiff’s failure to demonstrate practical significance cannot preclude that plaintiff from relying on competent evidence of statistical significance to establish a *prima facie* case of disparate impact”).

143. *E.g., Waisome*, 948 F.2d at 1376 (“Social scientists consider a finding of two standard deviations significant, meaning there is about one chance in 20 that the explanation for a deviation could be random. . . .”); *Adams v. Ameritech Serv., Inc.*, 231 F.3d 414, 424 (7th Cir. 2000) (“Two standard deviations is normally enough to show that it is extremely unlikely (. . . less than a 5% probability) that the disparity is due to chance.”); *Magistrini v. One Hour Martinizing Dry Cleaning*, 180 F. Supp. 2d 584, 605 n.26 (D.N.J. 2002) (a “statistically significant . . . study shows that there is only 5% probability that an observed association is due to chance”); *cf. Giles v. Wyeth, Inc.*, 500 F. Supp. 2d 1048, 1056 (S.D. Ill. 2007) (“While [plaintiff] admits that a p -value of .15 is three times higher than what scientists generally consider statistically significant—that is, a p -value of .05 or lower—she maintains that this ‘represents 85% certainty, which meets any conceivable concept of preponderance of the evidence.’”).

144. John P.A. Ioannidis, *Why Most Clinical Research Is Not Useful*, 13 PLOS Med. E1002049 (2016), <https://doi.org/10.1371/journal.pmed.1002049>; Patrick E. Shrout & Joseph L. Rodgers, *Psychology, Science, and Knowledge Construction: Broadening Perspectives from the Replication Crisis*, 69 Ann. Rev. Psych. 487 (2018), <https://doi.org/10.1146/annurev-psych-122216-011845>. For a description in the legal literature, see Beerden, *supra* note 10.

fields—prominent manifestos to “retire statistical significance”¹⁴⁵ and to move to “a world beyond ‘ $p < 0.05$ ’” have appeared.¹⁴⁶ Various alternatives or supplements to p -values are available, and the American Statistical Association (ASA) has issued statements on the best use of p -values and other quantities.¹⁴⁷ Its 2016 statement emphasized that p -values can be used to indicate the degree to which data are incompatible with a specified statistical model, but that they are not the probability that the model is true; that the size of the difference rather than just statistical significance should be reported; and that scientific conclusions or business decisions should not be based only on whether a test yields a statistically significant result.¹⁴⁸

Tests or Interval Estimates?

How can a highly significant difference be practically insignificant? The reason is simple: p depends not only on the magnitude of the effect, but also on the

145. Valentin Amrhein et al., *Retire Statistical Significance*, 567 *Nature* 305 (2019), <https://doi.org/10.1038/d41586-019-00857-9> (statement endorsed by more than 200 statisticians and other scientists calling “for a stop to the use of P values in the conventional, dichotomous way—to decide whether a result refutes or supports a scientific hypothesis. . . . P values, intervals and other statistical measures all have their place, but it’s time for statistical significance to go.”).

146. In a special issue of *The American Statistician*, the executive director of the American Statistical Association (ASA) and other authors proclaimed that “it is time to stop using the term ‘statistically significant’ entirely.” Ronald L. Wasserstein et al., *Moving to a World Beyond “ $p < 0.05$,”* 73:sup1 *Am. Statistician* 1, 2 (2019), <https://doi.org/10.1080/0031305.2019.1583913> (adding that “[n]or should variants such as ‘significantly different,’ ‘ $p < 0.05$,’ and ‘nonsignificant’ survive, whether expressed in words, by asterisks in a table, or in some other way”). However, the recommendation was not “official ASA policy.” Karen Kafadar, Editorial, *Statistical Significance, P-values, and Replicability*, 15 *Annals Applied Stat.* 1081 (2021), <https://doi.org/10.1214/21-AOAS1500>. For context on the *Nature* and ASA statements, see Benjamini, *supra* note 10.

147. Surprisingly, when $p = 0.05$, it is not very probable that a repetition of the same study under identical conditions would succeed in producing results for which $p < 0.05$. *E.g.*, Leonhard Held et al., *Replication Power and Regression to the Mean*, *Significance*, Dec. 2020, at 10–11 (arguing that this shows “the need for more stringent p -value thresholds to trust (original) ‘out-of-the-blue’ findings”); Laura C. Lazzeroni et al., *Solutions for Quantifying P-value Uncertainty and Replication Power*, 13 *Nature Methods* 107 (2016), <https://doi.org/10.1038/nmeth.3741>.

148. Ronald L. Wasserstein & Nicole A. Lazar, *ASA Statement on Statistical Significance and P-values*, 70 *Am. Statistician* 129 (2016), <https://dx.doi.org/10.1080/00031305.2016.1154108> (also noting in ¶ 3.6 that “a p -value near 0.05 taken by itself offers only weak evidence”). A task force appointed by the president of the American Statistical Association ecumenically concluded that “P-values, confidence intervals and prediction intervals [as well as] Bayes factors, posterior probability distributions and credible intervals are . . . some among many statistical methods useful for reflecting uncertainty” and that “P-values and significance tests, *when properly applied and interpreted*, increase the rigor of the conclusions drawn from data.” Yoav Benjamini et al., *ASA President’s Task Force Statement on Statistical Significance and Replicability*, 15 *Annals Applied Stat.* 1084 (2021), <https://doi.org/10.1214/21-AOAS1501> (emphasis added). On Bayes factors and the like, see section titled “Bayesian Statistical Methods and Posterior Probabilities” below.

sample size (among other things). With a huge sample, even a tiny effect will be highly significant.¹⁴⁹ For example, suppose that a company hires 52% of male job applicants and 49% of female applicants. With a large enough sample, a statistician could compute an impressively small p -value. This p -value would confirm that the difference does not result from chance, but it would not convert a minor difference (52% versus 49%) into a substantial one.¹⁵⁰ In short, the p -value does not measure the strength or importance of an association.

A “significant” effect can be small. Conversely, an effect that is “not significant” can be large. By inquiring into the magnitude of an effect, courts can avoid being misled by p -values. To focus attention on more substantive concerns—the size of the effect and the precision of the statistical analysis—interval estimates (e.g., confidence intervals) may be more valuable than tests. Seeing a plausible range of values for the quantity of interest helps describe the statistical uncertainty in the estimate.

Is the Sample Statistically Significant?

Many a sample has been praised for its statistical significance or blamed for its lack thereof. Technically, this makes little sense. Statistical significance is about the difference between observations and expectations. Significance therefore applies to statistics computed from the sample, but not to the sample itself, and certainly not to the size of the sample. Findings can be statistically significant. Differences can be statistically significant (see section titled “Is a Difference Statistically Significant?” above). Estimates can be statistically significant (see section titled “Statistical Models” below). By contrast, samples can be representative or unrepresentative. They can be chosen well or badly (see section titled “What Method Is Used to Select the Units?” above). They can be large enough to give reliable results or too small to bother with (see section titled “What is the Confidence Interval?” above). But samples cannot be “statistically significant,” if this technical phrase is to be used as statisticians use it.

149. See section titled “Is a Difference Statistically Significant?” above. Although some opinions seem to equate small p -values with “gross” or “substantial” disparities, most courts recognize the need to decide whether the underlying sample statistics reveal that a disparity is large. *E.g.*, *Washington v. People*, 186 P.3d 594 (Colo. 2008) (jury selection).

150. *Cf. Frazier v. Garrison Indep. Sch. Dist.*, 980 F.2d 1514, 1526 (5th Cir. 1993) (rejecting claims of intentional discrimination in the use of a teacher competency examination that resulted in retention rates exceeding 95% for all groups); *Washington*, 186 P.2d 594 (although a jury selection practice that reduced the representation of “African-Americans [from] 7.7 percent of the population [to] 7.4 percent of the county’s jury panels produced a highly statistically significant disparity, the small degree of exclusion was not constitutionally significant”).

Evaluating Hypothesis Tests

What Is the Power of the Test?

The power of a statistical study needs to be considered to avoid mistaking the absence of evidence for an effect with evidence for the absence of the effect. When a p -value is high, findings are not significant, and the null hypothesis is not rejected. This could happen for at least two reasons: (1) the null hypothesis is true; or (2) the null is false—but, by chance, the data happened to be of the kind expected under the null. If the power of a statistical study is low, the second explanation may be plausible. Power is the chance that a statistical test will declare an effect when there is an effect to be declared.¹⁵¹ This chance depends on the size of the effect and the size of the sample. Discerning subtle differences requires large samples; small samples may fail to detect substantial differences.

When a study with low power fails to show a significant effect, the results may therefore be more fairly described as inconclusive rather than negative. The proof is weak because power is low. On the other hand, when studies have a good chance of detecting a meaningful association, failure to obtain significance can be persuasive evidence that there is nothing much to be found.¹⁵²

151. More precisely, power is the probability of rejecting the null hypothesis when the alternative hypothesis (see section titled “What Are the Rival Hypotheses?” below) is right. Typically, this probability will depend on the values of unknown parameters, as well as the preset significance level α . The power can be computed for any value of α and any choice of parameters satisfying the alternative hypothesis. Frequentist hypothesis testing keeps the risk of a false positive to a specified level (such as $\alpha = 5\%$) and then tries to maximize power.

Statisticians usually denote power by the Greek letter beta (β). However, some authors use β to denote the probability of *accepting* the null hypothesis when the alternative hypothesis is true; this usage is fairly standard in epidemiology. Accepting the null hypothesis when the alternative holds true is a false negative (also called a Type II error, a missed signal, or a false acceptance of the null hypothesis).

The chance of a false negative may be computed from the power. Some commentators have claimed that the cutoff for significance should be chosen to equalize the chance of a false positive and a false negative, on the ground that this criterion corresponds to the more-probable-than-not burden of proof. The argument is fallacious, because $1 - \alpha$ and β do not give the probabilities of the null and alternative hypotheses. See sections titled “What Is the p -value?” and “Is a Difference Statistically Significant?” above; David H. Kaye, *Hypothesis Testing in the Courtroom*, in *Contributions to the Theory and Application of Statistics: A Volume in Honor of Herbert Solomon* 331, 341–43 (Alan E. Gelfand ed., 1987).

152. Some formal procedures (meta-analysis) are available to aggregate results across studies. See, e.g., *In re Bextra & Celebrex Marketing Sales Practices & Prod. Liab. Litig.*, 524 F. Supp. 2d 1166, 1174, 1184 (N.D. Cal. 2007) (holding that “[a] meta-analysis of all available published and unpublished randomized clinical trials” of certain pain-relief medicine was admissible). In principle, the power of the collective results will be greater than the power of each study. However, these procedures have their own weakness. See, e.g., Richard A. Berk & David A. Freedman, *Statistical Assumptions as Empirical Commitments*, in *Punishment and Social Control: Essays in Honor of Sheldon Messinger* 235, 244–48 (T.G. Blomberg & S. Cohen eds., 2d ed. 2003); Michael Oakes,

What About Small Samples?

For simplicity, the examples of statistical inference discussed here (see sections titled “Estimation” and “ p -values, Significance Levels, and Hypothesis Tests” above) were based on large samples. Small samples also can provide useful information. Indeed, when confidence intervals and p -values can be computed, the interpretation is the same with small samples as with large ones.¹⁵³ The concern with small samples is not that they are beyond the ken of statistical theory, but that:

1. It is hard to validate the assumptions underlying any proposed statistical approach.
2. Because approximations based on the normal curve generally cannot be used, suitable confidence intervals may be difficult to compute for parameters of interest. Likewise, p -values may be difficult to compute for hypotheses of interest.¹⁵⁴
3. Small samples may be unreliable, with large standard errors, broad confidence intervals, and tests having low power.

One Tail or Two?

Significance testing assesses the fit of the data to the null hypothesis within a given statistical model. In assessing whether the data fit the model, there is a choice to be made about whether to use a one-tailed or two-tailed p -value. The terms refer to the outcomes that would be considered as evidence of a lack of fit. The issue is easily explained with an example. Suppose we toss a coin 1,000 times and get 532 heads. The null hypothesis to be tested asserts that the coin is fair. The expected number of heads is 500; the excess number of heads is 32. If the null is correct, the chance of getting 532 or more heads is 2.3%. That would be used in a

Statistical Inference: A Commentary for the Social and Behavioral Sciences (1986); Diana B. Petitti, *Meta-Analysis, Decision Analysis, and Cost-Effectiveness Analysis Methods for Quantitative Synthesis in Medicine* (2d ed. 2000).

153. Advocates sometimes contend that samples are “too small to allow for meaningful statistical analysis,” *United States v. New York City Bd. of Educ.*, 487 F. Supp. 2d 220, 229 (E.D.N.Y. 2007), and courts often look to the size of samples from earlier cases to determine whether the sample data before them are admissible or convincing. *Id.* at 230; *Timmerman v. U.S. Bank*, 483 F.3d 1106, 1116 n.4 (10th Cir. 2007). However, a meaningful statistical analysis yielding a significant result can be based on a small sample, and reliability does not depend on sample size alone (see section titled “What Is the Confidence Interval?” above and the section titled “What Are the Slope and Intercept?” below). Well-known small-sample techniques include the sign test and Fisher’s exact test. *E.g.*, Michael O. Finkelstein & Bruce Levin, *Statistics for Lawyers* 154–56, 339–41 (2d ed. 2001); see generally E.L. Lehmann & H.J.M. d’Abrera, *Nonparametrics* (2d ed. 2006).

154. With large samples, approximate inferences (based on the central limit theorem, for example) may be quite adequate. These approximations will not be satisfactory for small samples unless the sampled values (not just the sample means) are approximately normally distributed.

one-tailed test, whose p -value is 2.3%. This test would be appropriate if we are only concerned that the coin is biased towards heads and thus are concerned only if we see too many heads. To make a two-tailed test, the statistician also computes the chance of a deficiency of 32 or more heads (that is, of getting $500 - 32 = 468$ heads or fewer). This probability is also 2.3%. So the probability of either an excess as large or larger than what occurred or a comparable deficiency is $2.3\% + 2.3\% = 4.6\%$. This larger probability is the two-tailed p -value. This value would be appropriate if we are concerned that the coin may be biased *either* towards heads or tails. In many cases, a statistical test can be done either one-tailed or two-tailed; the two-tailed method often produces a p -value twice as big as the one-tailed method. Because small p -values are evidence against the null hypothesis, the one-tailed test seems to produce stronger evidence than its two-tailed counterpart. However, the advantage is largely illusory, as the example suggests.

Some experts have argued for one or the other type of test,¹⁵⁵ but a rigid rule is not required if significance levels are used as guidelines rather than as mechanical rules for statistical proof. One-tailed tests often make it easier to reach a threshold such as 5%, at least in terms of appearance. However, if we recognize that 5% is not a magic line, then the choice between one tail and two is less important—as long as the choice and its effect on the p -value are made explicit.

How Many Tests Have Been Done?

Repeated testing complicates the interpretation of significance levels. If enough comparisons are made, random error almost guarantees that some will yield “significant” findings, even when there is no real effect. To illustrate the point, consider the problem of deciding whether a coin is biased. The probability that a fair coin will produce 10 heads when tossed 10 times is $(1/2)^{10} = 1/1024$. Observing 10 heads in the first 10 tosses therefore would be strong evidence that the coin is biased. Nonetheless, if a fair coin is tossed a few thousand times, it is likely that at least one string of ten consecutive heads will appear. Ten heads in the first ten tosses means one thing; a run of ten heads somewhere along the way to a few thousand tosses means quite another. A test—looking for a run of ten heads—can be repeated too often.

Artifacts from multiple testing are commonplace. Because research that fails to uncover significance often is not published, reviews of the literature may

155. See, e.g., *Jones v. Novartis Pharm. Corp.*, 235 F. Supp. 3d 1244, 1288–89 (N.D. Ala. 2017); *United States v. State of Del.*, 93 Fair Empl. Prac. Cas. (BNA) 1248, 2004 WL 609331, *10 n.27 (D. Del. 2004). According to formal statistical theory, the choice between one tail or two can sometimes be made by considering the exact form of the alternative hypothesis (see section titled “What Are the Rival Hypotheses?” below). But see Freedman et al., *supra* note 14, at 547–50. One-tailed tests at the 5% level are viewed as weak evidence—no weaker standard is commonly used in the technical literature. One-tailed tests are also called one-sided (with no pejorative intent); two-tailed tests are two-sided.

produce a misleadingly large number of studies finding statistical significance.¹⁵⁶ The many other nonsignificant tests are not part of the literature; but they have been carried out, and like the omitted tails in the coin flipping example, they have implications for interpreting the studies that have been published. Thus, multiple testing is a factor in the failure of many scientific studies to replicate.¹⁵⁷

Even a single researcher may examine so many different relationships that a few will achieve statistical significance by mere happenstance. For example, the researcher could choose a range of different outcome variables or different explanatory variables. Almost any large dataset—even pages from a table of random digits—will contain some unusual pattern that can be uncovered by diligent search. Having detected the pattern, the analyst can perform a statistical test for it, blandly ignoring the search effort.¹⁵⁸ Statistical significance is bound to follow.¹⁵⁹

There are statistical methods for dealing with multiple looks at the data, which permit the calculation of meaningful *p*-values in certain cases.¹⁶⁰ In addition, alternative approaches that focus on controlling the false discovery rate (the fraction of significant results expected to be false positives) are gaining in popularity. However, no general solution is known for handling multiple comparisons, and the existing methods would be of little help in the typical case where analysts have tested and rejected a variety of models before arriving at the one considered the most satisfactory (see section titled “Correlation and Regression” below on regression models). In these situations, researchers should be disclosing their multiple comparisons,¹⁶¹

156. E.g., Philippa J. Easterbrook et al., *Publication Bias in Clinical Research*, 337 *Lancet* 867 (1991), [https://doi.org/10.1016/0140-6736\(91\)90201-7](https://doi.org/10.1016/0140-6736(91)90201-7); John P.A. Ioannidis, *Effect of the Statistical Significance of Results on the Time to Completion and Publication of Randomized Efficacy Trials*, 279 *JAMA* 281 (1998); Stuart J. Pocock et al., *Statistical Problems in the Reporting of Clinical Trials: A Survey of Three Medical Journals*, 317 *New Eng. J. Med.* 426 (1987), <https://doi.org/10.1056/NEJM198708133170706>.

157. See section titled “Tests or Interval Estimates?” above.

158. The practice has been denominated HARKing. Norbert L. Kerr, *HARKing: Hypothesizing After the Results Are Known*, 2 *Personality & Social Psych. Rev.* 196 (1998), https://doi.org/10.1207/s15327957pspr0203_4.

159. Searching for significance in this way has been called data mining, data dredging, data snooping, selection bias, and, more recently, *p*-hacking. Megan L. Head et al., *The Extent and Consequences of P-Hacking in Science*, 13 *PLOS Biology* e1002106 (2015), <https://doi.org/10.1371/journal.pbio.1002106>.

160. See, e.g., Sandrine Dudoit & Mark J. van der Laan, *Multiple Testing Procedures with Applications to Genomics* (2008); Kaye, *supra* note 121, at 61–63; Martin Krzywinski & Naomi Altman, *Points of Significance: Comparing Samples—Part II*, 11 *Nature Methods* 355 (2014); Pak C. Sham & Shaun M. Purcell, *Statistical Power and Significance Testing in Large-scale Genetic Studies*, 15 *Nature Reviews Genetics* 335 (2014). In *Karlo v. Pittsburgh Glass Works*, 849 F.3d 61 (3d Cir. 2017), the court of appeals held that *Daubert* does not require an expert to use a particular (and conservative) adjustment method to find that a company’s reduction in force had a statistically significant disparate impact in three out of five overlapping subgroups of older workers. *Id.* at 83. For discussion of the case, see David H. Kaye, *Multiple Hypothesis Testing in Karlo v. Pittsburgh Glass Works*, *Forensic Sci., Stat.* & L., July 5, 2017, <https://perma.cc/JTD5-FULS>.

161. ASA Comm. on Professional Ethics, *supra* note 11, at 3.

and courts should not be overly impressed with claims that estimates are significant. Instead, they should be asking how analysts developed their models. Intuition may suggest that the more variables included in the model, the better. However, this idea often turns out to be wrong. Complex models and more variables may increase the chance of a spurious result (one that reflects only accidental features of the data). Standard statistical tests offer little protection against this possibility when the analyst has tried a variety of models before settling on the final specification.

What Are the Rival Hypotheses?

The p -value of a statistical test is computed on the basis of a model for the data: the null hypothesis. Usually, the test is made in order to argue for the alternative hypothesis: another model. However, on closer examination, both models may be unreasonable. A small p -value means something is going on besides random error. The alternative hypothesis should be viewed as one possible explanation, out of many, for the data.

In *Mapes Casino, Inc. v. Maryland Casualty Co.*,¹⁶² the court recognized the importance of explanations that the proponent of the statistical evidence had failed to consider. In this action to collect on an insurance policy, Mapes sought to quantify its loss from theft. It argued that employees were using an intermediary to cash in chips at other casinos. The casino established that over an 18-month period, the win percentage at its craps tables was 6%, compared to an expected value of 20%. The statistics proved that *something* was wrong at the craps tables—the discrepancy was too big to explain as the product of random chance. But the court was not convinced by plaintiff’s alternative hypothesis. The court pointed to other possible explanations (Runyonesque activities such as skimming, scamming, and cross-roading) that might have accounted for the discrepancy without implicating the suspect employees.¹⁶³ Rejection of the null hypothesis did not leave the proffered alternative hypothesis as the only viable explanation for the data.¹⁶⁴

162. 290 F. Supp. 186 (D. Nev. 1968).

163. *Id.* at 193. Skimming consists of “taking off the top before counting the drop,” scamming is “cheating by collusion between dealer and player,” and crossroading involves “professional cheaters among the players.” *Id.* In plainer language, the court seems to have ruled that the casino itself might be cheating, or there could have been cheaters other than the particular employees identified in the case. At the least, plaintiff’s statistical evidence did not rule out such possibilities. Compare *EEOC v. Sears, Roebuck & Co.*, 839 F.2d 302, 312 & n.9, 313 (7th Cir. 1988) (EEOC’s regression studies showing significant differences did not establish liability because surveys and testimony supported the rival hypothesis that women generally had less interest in commission sales positions), with *EEOC v. Gen. Tel. Co.*, 885 F.2d 575 (9th Cir. 1989) (unsubstantiated rival hypothesis of “lack of interest” in “nontraditional” jobs insufficient to rebut prima facie case of gender discrimination); cf. section titled “Is the Study Designed to Investigate Causation?” above (problem of confounding).

164. *E.g.*, *Coleman v. Quaker Oats Co.*, 232 F.3d 1271, 1283 (9th Cir. 2000) (A disparity with a p -value of “3 in 100 billion” did not demonstrate age discrimination because “Quaker never

Bayesian Statistical Methods and Posterior Probabilities

Standard errors, p -values, and significance tests are common techniques for assessing the potential impact of random errors or chance events. These procedures rely on sample data and their use is justified primarily in terms of the operating characteristics of statistical procedures.¹⁶⁵ They are often referred to as frequentist procedures because of this reliance on properties of the procedures that would be observed in repeated samples. As indicated in previous sections, frequentist procedures do not allow for assertions regarding the probability that a particular hypothesis is correct or that a particular confidence interval contains the true parameter value, given the data. For example, a statistician may postulate that a coin is fair: There is a 50–50 chance of landing heads, and successive tosses are independent. This is an empirical statement—potentially falsifiable—about the coin. It is easy to calculate the chance that a fair coin will turn up heads in every one of the next 10 tosses: The answer is $(1/2)^{10} = 1/1024$. Therefore, observing 10 heads in a row brings into serious doubt the initial hypothesis of fairness.

But what of the converse probability: If the coin does land heads 10 times, what is the chance that it is fair?¹⁶⁶ To compute such converse probabilities, it is necessary to postulate initial probabilities that the coin is fair, as well as probabilities of unfairness to various degrees. In the Bayesian approach, probabilities represent subjective degrees of belief about hypotheses or causes rather than objective facts about observations. The observer must quantify beliefs about the chance that the coin is unfair to various degrees—in advance of seeing the data. For example, let p be the unknown probability that the coin lands heads. What is the chance that p exceeds 0.1? 0.6? These subjective probabilities, like the probabilities governing the tosses of the coin, are set up to obey the axioms of probability

contends that the disparity occurred by chance; just that it did not occur for discriminatory reasons. When other pertinent variables were factored in, the statistical disparity diminished and finally disappeared.”).

165. Operating characteristics include the expected value and standard error of estimators, probabilities of error for statistical tests, and the like.

166. We call this a converse probability because it is of the form $P(H_0|\text{data})$ rather than $P(\text{data}|H_0)$; an equivalent phrase, “inverse probability,” also is used. Treating $P(\text{data}|H_0)$ as if it were the converse probability $P(H_0|\text{data})$ is the transposition fallacy. For example, most U.S. senators are men, but few men are senators. (As of 2024, 75 of the 100 senators are men.) Consequently, there is a high probability (0.75) that an individual who is a senator is a man, but the probability that an individual who is a man is a senator is practically zero. For examples of the transposition fallacy in court opinions, see cases cited *supra* notes 126 & 136. The frequentist p -value, $P(\text{data}|H_0)$, is generally not a good approximation to the Bayesian $P(H_0|\text{data})$; the latter includes considerations of power and base rates.

theory.¹⁶⁷ The probabilities for the various hypotheses about the coin, specified before data collection, are called prior probabilities. Prior probabilities can be updated, using Bayes' rule, given data on how the coin actually falls. (The Appendix explains the rule.) In short, a Bayesian analysis involves posterior probabilities for various hypotheses about the coin, given the data. These posterior probabilities quantify the statistician's confidence in the hypothesis that a coin is fair.¹⁶⁸ Although such posterior probabilities relate directly to hypotheses of legal interest, they are necessarily subjective, for they reflect not just the data but also the subjective prior probabilities—that is, degrees of belief about hypotheses formulated prior to obtaining data.¹⁶⁹

With the exceptions of parentage testing and so-called probabilistic genotyping software for interpreting complex DNA mixtures,¹⁷⁰ such analyses have rarely been used in court, and the question of their forensic value was aired primarily in the academic literature. Some statisticians favor Bayesian methods, and such methods have become increasingly popular for settings in which a number of parameters are to be estimated simultaneously (for example, in assessing the effectiveness of standardized test coaching programs offered at a number of different schools).¹⁷¹ Some legal commentators have proposed using these methods in various settings.¹⁷²

167. Bayesian procedures are sometimes defended on the ground that the beliefs of any rational observer must conform to the Bayesian rules. However, the definition of “rational” is purely formal. See Peter C. Fishburn, *The Axioms of Subjective Probability*, 1 Stat. Sci. 335 (1986); David Freedman, *Some Issues in the Foundation of Statistics*, 1 Found. Sci. 19 (1995), reprinted in *Topics in the Foundation of Statistics* 19 (Bas C. van Fraassen ed., 1997); David Kaye, *The Laws of Probability and the Law of the Land*, 47 U. Chi. L. Rev. 34 (1979).

168. Here, confidence has the meaning ordinarily ascribed to it, rather than the technical interpretation applicable to a frequentist confidence interval. Consequently, it can be related to the burden of persuasion. See David H. Kaye, *Apples and Oranges: Confidence Coefficients and the Burden of Persuasion*, 73 Cornell L. Rev. 54 (1987).

169. But see *infra* note 226.

170. See Kaye, *supra* note 55.

171. Donald B. Rubin, *Estimation in Parallel Randomized Experiments*, 6 J. Educ. Stat. 377 (1981), <https://doi.org/10.3102/10769986006004377>. Many practicing statisticians are pragmatists, using whatever procedure they think is appropriate for the occasion.

172. See Christopher S. Elmendorf & Douglas M. Spencer, *Administering Section 2 of the Voting Rights Act After Shelby County*, 115 Colum. L. Rev. 2143, 2215 (2015); David H. Kaye, *Forensic Statistics in the Courtroom*, in *Handbook of Forensic Statistics* 225–48 (David Banks et al. eds., 2021); Kaye et al., *supra* note 8, §§ 12.8.5 & 14.3.2; David H. Kaye, *Rounding Up the Usual Suspects: A Legal and Logical Analysis of DNA Database Trawls*, 87 N.C. L. Rev. 425 (2009). In addition, as indicated in the Appendix, Bayes' rule is crucial in solving certain problems involving conditional probabilities of related events. For example, if the proportion of women with breast cancer in a region is known, along with the probability that a mammogram of an affected woman will be positive for cancer and that the mammogram of an unaffected woman will be negative, then one can compute the numbers of false-positive and false-negative mammography results that would be expected to arise in a population-wide screening program. Using Bayes' rule to diagnose a specific patient, however, is more problematic, because the prior probability that the patient has breast cancer may

Correlation and Regression

Regression models are used by many social scientists to infer causation from association. Such models have been offered in court to prove disparate impact in discrimination cases, to estimate damages in antitrust actions, and for many other purposes. The sections titled “Scatter Diagrams,” “Correlation Coefficients,” and “Regression Lines” cover some preliminary material, showing how scatter diagrams, correlation coefficients, and regression lines can be used to summarize relationships between variables.¹⁷³ The section titled “Statistical Models” explains the ideas of regression modeling and some of the pitfalls, particularly those arising in the use of regression models to infer causation.

Scatter Diagrams

The relationship between two variables can be graphed in a scatter diagram (also called a scatterplot or scattergram). We begin with data on income and education for a sample of 238 men, ages 25 to 34, residing in Kansas.¹⁷⁴ Each person in the sample corresponds to one dot in the diagram. As indicated in Figure 5, the horizontal axis shows education, and the vertical axis shows income. Person A completed 12 years of schooling (high school) and had an income of \$50,000. Person B completed 16 years of schooling (college) and had an income of \$100,000.

not equal the population proportion. Nevertheless, to overcome the tendency to focus on a test result without considering the “base rate” at which a condition occurs, a diagnostician can apply Bayes’ rule to plausible base rates before making a diagnosis. Finally, Bayes’ rule also is valuable as a device to explicate the meaning of concepts such as error rates, probative value, and transposition. See, e.g., David H. Kaye, *The Double Helix and the Law of Evidence* (2010); Kaye et al., *Wigmore*, *supra* note 44, § 7.3.2; David H. Kaye, *Digging into the Foundations of Evidence Law*, 115 Mich. L. Rev. 915 (2017).

173. The focus is on simple linear regression. See also Rubinfeld & Card, *supra* note 25, for further discussion of these ideas with an emphasis on econometrics.

174. These data are from a public-use data file of the 2021 American Community Survey and were obtained from the data.census.gov website of the Bureau of the Census, U.S. Department of Commerce (precise query: <https://data.census.gov/mdat/#/search?ds=ACSPUMS1Y2021&vv=PERNP,%2aAGEP&cv=SEX&rv=ucgid,SCHL&wt=PWGTP&g=0400000US20>). Income and education are self-reported. Income is censored at \$200,000. Only positive income values are included. Education variables are recoded to yield years of education. Data are a 20% sample from the resulting dataset to make the figures easier to read. Both variables in a scatter diagram have to be quantitative (with numerical values) rather than qualitative (nonnumerical).

Figure 5. Plotting points in a scatter diagram.

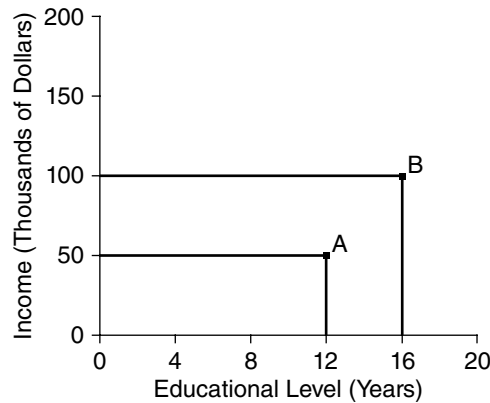
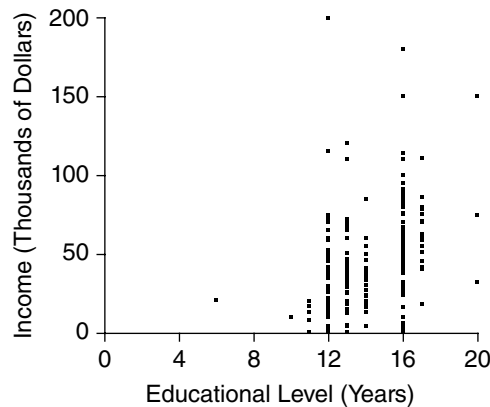


Figure 6 is the scatter diagram for the Kansas data. The diagram confirms an obvious point: There is a positive association between income and education. In general, persons with a higher educational level have higher incomes. However, there are many exceptions to this rule, and the association is not as strong as one might expect.

Figure 6. Scatter diagram for income and education: men ages 25 to 34 in Kansas.

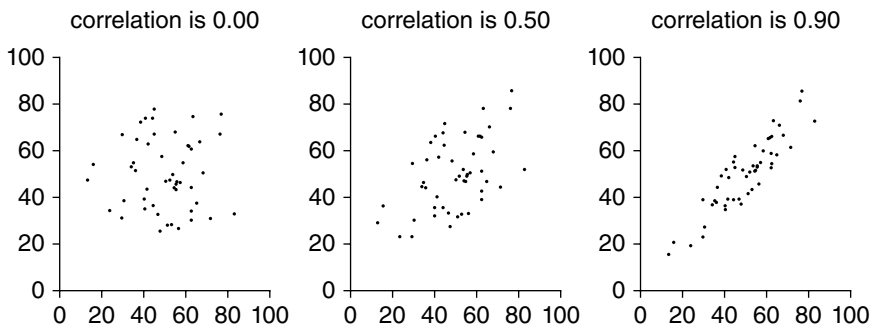


Correlation Coefficients

Two variables are positively correlated when their values tend to go up or down together, such as income and education in Figure 6. The correlation coefficient

(usually denoted by the letter r) is a single number that reflects the magnitude of a linear association and whether it is positive or negative. Figure 7 shows r for three scatter diagrams: In the first, there is no association; in the second, the association is positive and moderate; in the third, the association is positive and strong. Moving across these diagrams, the clouds of dots are increasingly clustered around a straight line that could be drawn through the cloud.

Figure 7. The correlation coefficient measures the sign of a linear association and its strength.



A correlation coefficient of 0 indicates no linear association between the variables. The maximum value for the coefficient is +1, indicating that all the dots fall exactly on a straight line that slopes up. Sometimes, there is a negative association between two variables: Large values of one tend to go with small values of the other. The age of a car and its fuel economy in miles per gallon illustrate the idea. Negative association is indicated by negative values for r . The extreme case is an r of -1 , indicating that all the points in the scatter diagram lie on a straight line that slopes down.

Weak associations are the rule in the social sciences. In Figure 6, the correlation between income and education is about 0.4. The correlation between college grades and first-year law school grades is under 0.3 at most law schools, while the correlation between LSAT scores and first-year grades is generally about 0.4.¹⁷⁵ The correlation between heights of fraternal twins is roughly 0.5. The correlation between heights of identical twins is about 0.9.¹⁷⁶

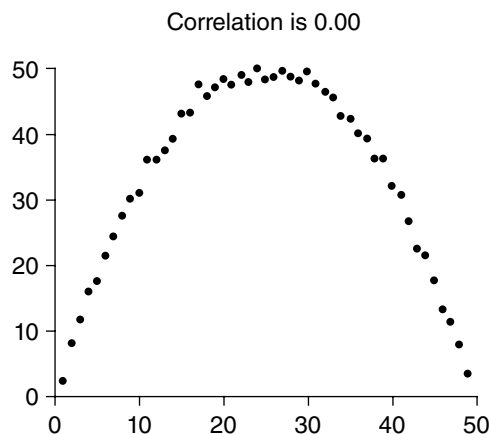
175. Law School Admissions Council, Summary of 2017, 2018, and 2019 LSAT Correlation Study Results, <https://perma.cc/452D-JVNP>. Adjusting for range restriction boosts the median correlations to 0.44 and 0.61. *Id.*

176. G. Frederiek Estourgie-van Burk et al., *Body Size in Five-year-old Twins: Heritability and Comparison to Singleton Standards*, 9 *Twin Research & Hum. Genetics* 646, 649 tbl. 2 (2006), <https://doi.org/10.1375/183242706778553417> (reporting correlations of about 0.59 and 0.94 for the two types of

Is the Association Linear?

The correlation coefficient has a number of limitations, to be considered in turn. The correlation coefficient is designed to measure linear association. Figure 8 shows a strong nonlinear pattern with a correlation close to zero. The correlation coefficient is of limited use with a nonlinear relationship.

Figure 8. A strong nonlinear association with a correlation coefficient close to zero. The correlation coefficient only measures the degree of linear association.



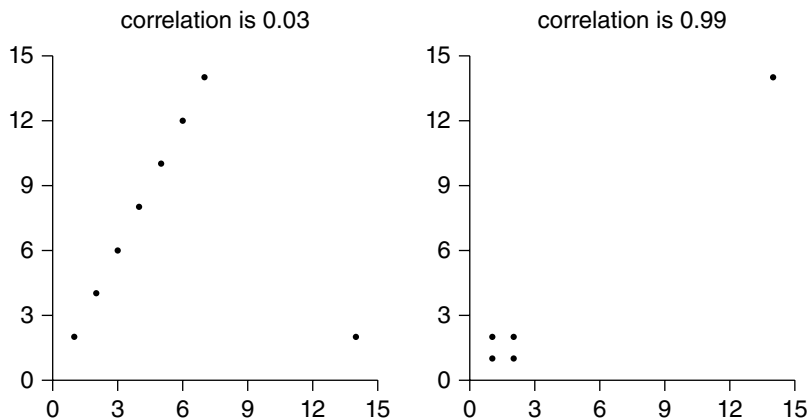
Do Outliers Influence the Correlation Coefficient?

The correlation coefficient can be distorted by outliers—a few points that are far removed from the bulk of the data. The left-hand panel in Figure 9 shows that one outlier (lower right-hand corner) can reduce a perfect correlation to nearly nothing. Conversely, the right-hand panel shows that one outlier (upper right-hand corner) can raise a correlation of zero to nearly one.¹⁷⁷ If there are extreme outliers in the data, the correlation coefficient is unlikely to be meaningful.

twins); Matthew C. Keller et al., *The Genetic Correlation Between Height and IQ: Shared Genes or Assortative Mating?*, 10 PLOS Genetics e1004329 tbl. 2 (2013), <https://doi.org/10.1371/journal.pgen.1003451> (0.46 and 0.87); Jon Martin Sundet et al., *Resolving the Genetic and Environmental Sources of the Correlation Between Height and Intelligence: A Study of Nearly 2600 Norwegian Male Twin Pairs*, 8 Twin Research & Hum. Genetics 307, 309 tbl. 2 (2005), <https://doi.org/10.1375/1832427054936745> (0.54 and 0.92).

177. Cf. David H. Kaye, *The Dynamics of Daubert: Methodology, Conclusions, and Fit in Statistical and Econometric Studies*, 87 Va. L. Rev 1933 (2001) (describing a simple linear regression in an anti-trust case in which “[t]he difference between a finding by [plaintiff’s expert] of several hundred

Figure 9. The correlation coefficient can be distorted by outliers.



Does a Confounding Variable Influence the Coefficient?

The correlation coefficient measures the association between two variables. Researchers—and the courts—are usually more interested in causation. Causation is not the same as association. The association between two variables may be driven by a lurking variable that has been omitted from the analysis (see section titled “Is the Study Designed to Investigate Causation?” above). For an easy example, there is an association between shoe size and vocabulary among school-children. However, learning more words does not cause the feet to get bigger, and larger feet do not make children more articulate. In this case, the lurking variable is easy to spot—age. In more realistic examples, the lurking variable may be harder to identify.

In statistics, lurking variables are called confounders or confounding variables. Association may reflect causation, but a large correlation coefficient is not enough to warrant causal inference. A large value of r means only that the dependent variable marches in step with the independent one: Possible reasons include causation, confounding, and coincidence. Multiple regression is one method that attempts to deal with confounders (see “Statistical Models” below).¹⁷⁸

million dollars of damages and a finding of no damages is the inclusion in his model of a single anomalous data point”—a result that “would not be acceptable in an undergraduate econometrics class, let alone professional work” but that went undetected at trial) (quoting Brief of Amicus Curiae Dr. Daniel L. McFadden in Support of Defendants-Appellants at 15, 19, *Conwood Co. v. United States Tobacco Co.*, 290 F.3d 768 (2002)).

178. See also Rubinfeld & Card, *supra* note 25. The difference between experiments and observational studies is discussed in the section titled “Descriptive Surveys and Censuses” above.

Regression Lines

The regression line can be used to describe a linear trend in the data. The regression line for income on education in the Kansas sample is shown in Figure 10. The height of the line estimates the average income for a given educational level. For example, the average income for people with 12 years of education is estimated at \$35,600, indicated by the height of the line at 12 years. The average income for people with 16 years of education is estimated at \$58,400.

Figure 10. The regression line for income on education and its estimates.

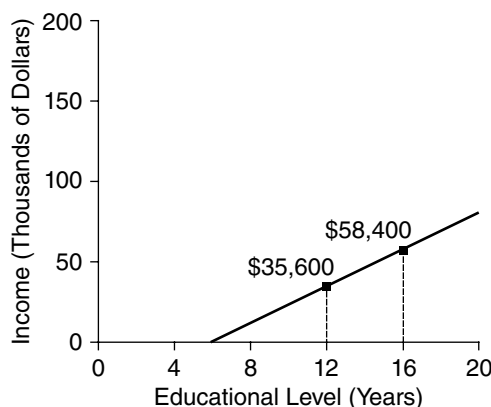
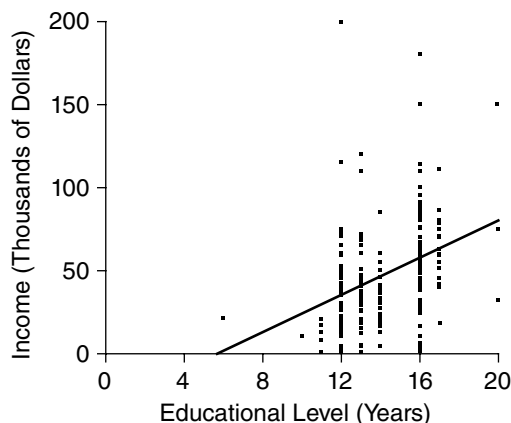


Figure 11 shows the points for income and education that were plotted in Figure 6 with the regression line superimposed. Many straight lines could have been drawn through the data points. Some would “fit” the data better than others. The regression line in Figure 11 comes from a method known as “ordinary least squares” (OLS). Every point in the scatterplot lies some distance directly above or below the line. We can square these distances and add them up. The OLS regression line is the one line for which this sum of the squared distances is the smallest. It shows the average trend of income as education increases. Thus, the regression line indicates the extent to which a change in one variable (income) is associated with a change in another variable (education).

What Are the Slope and Intercept?

The regression line can be described in terms of its intercept and slope. Often, the slope is the more interesting statistic. In Figure 10, the slope is \$5,700 per year. On average, each additional year of education is associated with an additional \$5,700 of income. Next, the intercept is $-\$32,800$. This is an estimate of

Figure 11. Scatter diagram for income and education, with the regression line indicating the trend.



the average income for the (hypothetical) population of persons with zero years of education.¹⁷⁹ In this case, the estimate is nonsensical (below zero!) because zero years of education is very far from the bulk of the data where income appears to grow linearly with increasing years of education.

The slope of the regression line has the same limitations as the correlation coefficient: (1) The slope may be misleading if the relationship is strongly non-linear; and (2) the slope may be affected by confounders and outliers. With respect to (1), the slope of \$5,700 per year in Figure 10 presents each additional year of education as having the same value, but some years of schooling surely are worth more and others less. With respect to (2), the association between education and income is no doubt causal, but there are other factors to consider, including family background. Compared to individuals who did not graduate from high school, people with college degrees usually come from richer and better-educated families. Thus, college graduates have advantages besides education. As statisticians

179. The regression line, like any straight line, has an equation of the form $y = a + bx$. Here, a is the intercept (the value of y when $x = 0$), and b is the slope (the change in y per unit change in x). In Figure 11, the intercept of the regression line is $-\$32,800$ and the slope is $\$5,700$ per year. The line estimates an average income of $\$58,400$ for people with 16 years of education. This may be computed from the intercept and slope as follows:

$$-\$32,800 + (\$5,700 \text{ per year}) \times 16 \text{ years} = -\$32,800 + \$91,200 = \$58,400.$$

The slope b is the same anywhere along the line. Mathematically, that is what distinguishes straight lines from other curves. If the association is negative, the slope will be negative too. The slope is like the grade of a road, and it is negative if the road goes downhill. The intercept is like the starting elevation of a road, and it is computed from the data so that the line goes through the center of the scatter diagram, rather than being generally too high or too low.

might say, the effects of family background are confounded with the effects of education. Statisticians often use the guarded phrases “on average” and “associated with” when talking about the slope of the regression line. This is because the slope has limited utility when it comes to making causal inferences.

What Is the Unit of Analysis?

If association between characteristics of individuals is of interest, these characteristics should be measured on individuals. Sometimes individual-level data do not exist, but rates or averages for groups are available. “Ecological” correlations are computed from such rates or averages. These correlations generally overstate the strength of an association. For example, average income and average education can be determined for men living in each state and in Washington, D.C. The correlation coefficient for these 51 pairs of averages turns out to be 0.7. However, states do not go to school and do not earn incomes. People do. The correlation for income and education for men in the United States is only 0.4. The correlation for state averages overstates the correlation for individuals—a common tendency for ecological correlations.¹⁸⁰

Ecological analysis is often seen in cases claiming dilution in voting strength of minorities. In this type of voting rights case, plaintiffs must prove three things: (1) the minority group constitutes a majority in at least one district of a proposed plan; (2) the minority group is politically cohesive—that is, votes fairly solidly for its preferred candidate; and (3) the majority group votes sufficiently as a bloc to defeat the minority-preferred candidate.¹⁸¹ The first requirement is compactness; the second and third define polarized voting.

180. Correlations are computed from the March 2005 Current Population Survey for men ages 25–64. Freedman et al., *supra* note 14, at 149. The ecological correlation uses only the average figures, but within each state there is a lot of spread about the average. The ecological correlation smoothes away this individual variation. *Cf.* Gold et al., *supra* note 15 (suggesting that ecological studies of exposure and disease are “far from conclusive” because of the lack of data on confounding variables (a much more general problem) as well as the possible aggregation bias described here); David A. Freedman, *Ecological Inference and the Ecological Fallacy*, in 6 *Int’l Encyclopedia of the Social and Behavioral Sciences* 4027 (Neil J. Smelser & Paul B. Baltes eds., 2001).

181. See *Thornburg v. Gingles*, 478 U.S. 30, 50–51 (1986) (“First, the minority group must be able to demonstrate that it is sufficiently large and geographically compact to constitute a majority in a single-member district. . . . Second, the minority group must be able to show that it is politically cohesive. . . . Third, the minority must be able to demonstrate that the white majority votes sufficiently as a bloc to enable it . . . usually to defeat the minority’s preferred candidate.”). In subsequent cases, the Court has emphasized that these factors are not sufficient to make out a violation of section 2 of the Voting Rights Act. *E.g.*, *Johnson v. De Grandy*, 512 U.S. 997, 1011 (1994) (“*Gingles* . . . clearly declined to hold [these factors] sufficient in combination, either in the sense that a court’s examination of relevant circumstances was complete once the three factors were found to exist, or in the sense that the three in combination necessarily and in all circumstances

The secrecy of the ballot box means that polarized voting cannot be directly observed. Instead, plaintiffs in voting rights cases rely on ecological regression, with scatter diagrams, correlations, and regression lines to estimate voting behavior by groups and demonstrate polarization.¹⁸² The unit of analysis typically is the precinct. For each precinct, public records can be used to determine the percentage of registrants in each demographic group of interest, as well as the percentage of the total vote for each candidate—by voters from all demographic groups combined. Plaintiffs’ burden is to determine the vote by each demographic group separately.

Figure 12 shows how the argument unfolds. Each point in the scatter diagram represents data for one precinct in the 1982 Democratic primary election for auditor in Lee County, South Carolina. The horizontal axis shows the percentage of registrants who are white. The vertical axis shows the turnout rate for the white candidate. The regression line is plotted too. The slope would be interpreted as the expected increase in the proportion supporting the white candidate for each 1% increase in the proportion of white registrants in the precinct. The intercept then would be the expected support for the white candidate in a precinct with no white registrants (an all-Black precinct).¹⁸³ The validity of such estimates is contested in the statistical and legal literature.¹⁸⁴

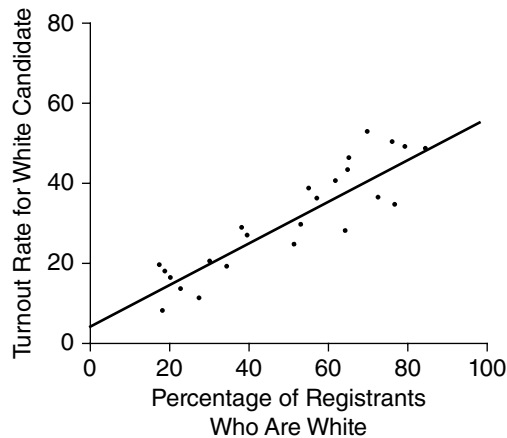
demonstrated dilution.”). On the legal framework and its problems, see Christopher S. Elmendorf et al., *Racially Polarized Voting*, 83 U. Chi. L. Rev. 587 (2016); D. James Greiner, *Re-Solidifying Racial Bloc Voting: Empirics and Legal Doctrine in the Melting Pot*, 86 Ind. L.J. 447 (2011); Nicholas O. Stephanopoulos, *The Relegation of Polarization*, 83 U. Chi. L. Rev. Online 160, 168 (2017).

182. Less readily visualized procedures also are used. See, e.g., *Ecological Inference: New Methodological Strategies* (Gary King et al. eds., 2004); Loren Collingwood et al., *The R Journal: eiCompare: Comparing Ecological Inference Estimates Across EI and ELRxC*, R J., Dec. 2016, at 92, <https://perma.cc/U5JH-GBEM>. We confine our explanation to the simplest (bivariate linear) form of ecological regression. All the methods confront the same fundamental problem of inferring associations at the level of individuals from differences among groups that include heterogeneous subgroups of individuals. Given this common issue and objective, one might think that the phrase “ecological inference” would be a broad umbrella term, but especially in voting rights litigation, the phrase usually refers to one particular procedure that is more complex than traditional bivariate ecological regression. See, e.g., *Luna v. County of Kern*, 291 F. Supp. 3d 1088, 1118 (E.D. Cal. 2018) (describing differences between ecological regression (ER) and “[e]cological inference (‘EI’), developed by political scientist Gary King in 1997 [that] is ‘similar to, but largely regarded as an improvement upon’ the ER methodology endorsed in *Gingles*”). But see Elmendorf & Spencer, *supra* note 172, at 2180 (referring to “statistically tenuous techniques of ecological inference”); Greiner, *supra* note 181, at 449 (observing that “analyzing census data (usually) and precinct-level vote returns via a set of methods collectively called ‘ecological inference’ has been a staple since the mid 1980s”).

183. By definition, the turnout rate equals the number of votes for the candidate, divided by the number of registrants; the rate is computed separately for each precinct. The intercept of the line in Figure 11 is 4%, and the slope is 0.52. Plaintiffs would conclude that only 4% of the voters in an all-Black precinct would be expected to vote for the white candidate, while $4\% + 52\% = 56\%$ of the voters in an all-white precinct would be expected to vote for the white candidate, which demonstrates polarization.

184. E.g., Moon Duchin & Douglas M. Spencer, *Models, Race, and the Law*, 130 Yale L.J. Forum 744 (2021); Elmendorf & Spencer, *supra* note 172, at 2203 (“Ecological inference as

Figure 12. Turnout rate for the white candidate plotted against the percentage of registrants who are white. Precinct-level data, 1982 Democratic primary for auditor, Lee County, South Carolina.¹⁸⁵



Statistical Models

Statistical models are widely used in the social sciences and in litigation. For example, the census suffers an undercount, more severe in certain places than others.

traditionally practiced relies on heroic assumptions, elides questions about statistical precision, and uses post-hoc corrections to paper over mathematically impossible results. . . .”) (note omitted). For references to some of the statistical literature, see Collingwood et al., *supra* note 182, at 92–94. The use of ecological regression and other quantitative methods increased considerably after the Supreme Court noted in *Thornburg v. Gingles*, 478 U.S. 30, 53 n.20 (1986), that “[t]he District Court found both methods [extreme case analysis and bivariate ecological regression analysis] standard in the literature for the analysis of racially polarized voting.” See Bruce M. Clarke & Robert Timothy Reagan, Federal Judicial Center, *Redistricting Litigation: An Overview of Legal, Statistical, and Case-Management Issues* (2002); D. James Greiner, *The Quantitative Empirics of Redistricting Litigation: Knowledge, Threats to Knowledge, and the Need for Less Districting*, 29 Yale L. & Policy Rev. 527 (2011); D. James Greiner, *Ecological Inference in Voting Rights Act Disputes: Where Are We Now, and Where Do We Want to Be?*, 47 Jurimetrics J. 115, 117, 121 (2007).

Redistricting plans based predominantly on racial considerations are unconstitutional unless narrowly tailored to meet a compelling state interest. *Shaw v. Reno*, 509 U.S. 630 (1993). Whether compliance with the Voting Rights Act can be considered a compelling interest is an open question, but efforts to sustain racially motivated redistricting on this basis have not fared well before the Supreme Court. See *Abrams v. Johnson*, 521 U.S. 74 (1997); *Shaw v. Hunt*, 517 U.S. 899 (1996); *Bush v. Vera*, 517 U.S. 952 (1996); *Abbott v. Perez*, 585 U.S. 579 (2018); *but see Alabama Legislative Black Caucus v. Alabama*, 575 U.S. 254 (2015); *Bethune-Hill v. Virginia State Bd. of Elections*, 580 U.S. 178 (2017); *Cooper v. Harris*, 581 U.S. 285 (2017).

185. Data from James W. Loewen & Bernard Grofman, *Recent Developments in Methods Used in Vote Dilution Litigation*, 21 Urb. Law. 589, 591 tbl.1 (1989).

If some statistical models are to be believed, the undercount can be corrected—moving seats in Congress and millions of dollars a year in tax funds.¹⁸⁶ Other models purport to lift the veil of secrecy from the ballot box, enabling the experts to determine how minority groups have voted—a crucial step in voting rights litigation (see section titled “Regression Lines” above). This section discusses the statistical logic of regression models.

A regression model attempts to combine the values of certain variables (the independent variables) to get expected values for another variable (the dependent variable). The model can be expressed in the form of a regression equation. A simple regression equation has only one independent variable; a multiple regression equation has several independent variables. Coefficients in the equation will be interpreted as showing the effects of changing the corresponding variables. This is justified in some situations, as the next example demonstrates.

Hooke’s law (named after Robert Hooke, England, 1653–1703) describes how a spring stretches in response to a load: Strain is proportional to stress. To verify Hooke’s law experimentally, a physicist will make a number of observations on a spring. For each observation, the physicist hangs a weight on the spring and measures its length. A statistician could develop a regression model for these data:

$$length = a + (b \times weight) + \varepsilon \quad (1)$$

The error term, denoted by the Greek letter ε (epsilon) is needed because measured length will not be exactly equal to $a + (b \times weight)$. If nothing else, error in measuring the spring’s length must be reckoned with.¹⁸⁷ The model takes ε as “random error”—behaving like draws made at random with replacement from a box of tickets. Each ticket shows a potential error, which will be realized if that ticket is drawn. The average of the potential errors in the box is assumed to be zero.¹⁸⁸

Equation (1) has two parameters, a and b . These constants of nature characterize the behavior of the spring: a is length under no load, and b is elasticity (the increase in length per unit increase in weight). By way of numerical illustration, suppose a is 400 and b is 0.05. If the weight is 1, the length of the spring is expected to be

$$400 + (0.05 \times 1) = 400.05.$$

186. See Brown et al., *supra* note 36.

187. We will assume that the physicist uses a perfectly accurate and unchanging set of standard weights.

188. In standard statistical terminology, the process of measuring length here is unbiased.

If the weight is 3, the expected length is

$$400 + (0.05 \times 3) = 400 + 0.15 = 400.15.$$

In either case, the actual length will differ from expected, by a random error ε .

In the simplest situation, the ε 's for different observations on the spring are assumed to be independent and identically distributed, with a mean of zero. "Independent" means that, for any two observations using the same weight, the chances for one ε do not depend on outcomes for the other. If the errors are like draws made at random with replacement from a box of tickets, as we assumed earlier, they are independent—the box will not change from one draw to the next. "Identically distributed" means that the chance behavior of the two ε 's is the same: They are drawn at random from the same box.

The parameters a and b in equation (1) are not directly observable, but they can be estimated by the method of least squares.¹⁸⁹ Statisticians often denote estimates by hats. Thus, $\hat{a}$ is the estimate for a , and $\hat{b}$ is the estimate for b . The values of $\hat{a}$ and $\hat{b}$ are chosen to minimize the sum of the squared prediction errors. These errors are also called residuals. They measure the difference between the actual length of the spring and the predicted length, the latter being $\hat{a} + \hat{b} \times \text{weight}$:

$$\text{actual length} = \hat{a} + (\hat{b} \times \text{weight}) + \text{residual} \quad (2)$$

Of course, no one really imagines there to be a box of tickets hidden in the spring. However, the variability of physical measurements (under many but by no means all circumstances) does seem to be remarkably like the variability in draws from a box.¹⁹⁰ In short, the statistical model corresponds rather closely to the empirical phenomenon.

Equation (1) is a statistical model for the data, with unknown parameters a and b . The error term ε is not observable. The model is a theory—and a good one—about how the data are generated. By contrast, equation (2) is a regression equation that is fitted to the data: The intercept $\hat{a}$, the slope $\hat{b}$, and the residual can all be computed from the data. The results are useful because $\hat{a}$ is a good estimate for a , and $\hat{b}$ is a good estimate for b . (Similarly, the residual is a good approximation

189. See section titled "Regression Lines" above. It might seem that a is observable; after all, we can measure the length of the spring with no load. However, the measurement is subject to error, so we observe not a , but $a + \varepsilon$. See equation (1). The parameters a and b can be estimated, even estimated very well, but they cannot be observed directly. The least squares estimates of a and b are the intercept and slope of the regression line. See section titled "What Are the Slope and Intercept?" above and Freedman et al., *supra* note 14, at 208–10. The method of least squares was developed by Adrien-Marie Legendre (France, 1752–1833) and Carl Friedrich Gauss (Germany, 1777–1855) to fit astronomical orbits.

190. This is the Gauss model for measurement error. See Freedman et al., *supra* note 14, at 450–52.

to ϵ .) Without the theory, these estimates would be less useful. Is there a theoretical model behind the data processing? Is the model justifiable? These questions can be critical when it comes to making statistical inferences from the data.

These points apply to more complicated statistical models. The simple linear regression model exemplified in equation (1) can be extended in many ways. If another variable might contribute to or be associated with a response, an additional term (*constant $\times$ variable*) can be added to the right-hand side of the equation. Furthermore, the relationship between the independent and dependent variables need not be linear. For instance, to model measurements of the distance d a ball dropped from the Leaning Tower of Pisa falls in a few seconds, Newton's laws imply that we should use a model with the time variable squared.¹⁹¹ Also, variables that are dichotomous rather than continuous (such as whether a mouse will develop cancer after exposure to a food additive) can be part of a regression model.¹⁹²

In social science and legal applications, such advanced statistical models often are invoked—even when they lack an independent theoretical basis. Hooke's law—incorporated in equation (1)—is relatively easy to test experimentally. For something like the salaries that an employer pays to workers, validation of a proposed relationship between the dependent variable and the independent ones would be difficult to validate experimentally. When expert testimony relies on statistical models, the court may well inquire, what are the assumptions behind the model, and why do they apply to the case at hand? The assumptions being questioned can include the choice of variables in the regression and the nature of the assumed relationship. It is important to distinguish between two situations:

- The nature of the relationship between the variables is known and regression is being used to make quantitative estimates of parameters in that relationship, and
- The nature of the relationship is largely unknown and regression is being used to determine the nature of the relationship—or indeed whether any relationship exists at all.

Regression was developed to handle situations of the first type, with Hooke's law being an example. The basis for the second type of application is analogical, and the tightness of the analogy is an issue worth exploration.¹⁹³

191. The model would be $distance = a + (b \times seconds^2) + \epsilon$ instead of $distance = a + (b \times seconds) + \epsilon$.

192. Such extensions and modifications of simple linear regression are introduced in the *Reference Guide on Multiple Regression and Advanced Statistical Models*; see also Andrew Gelman et al., *Regression and Other Stories* (2020).

193. See, e.g., David A. Freedman, *Statistical Models and Causal Inference: A Dialogue with the Social Sciences* (David Collier et al. eds., 2009); Andrew Gelman, *Review Essay: Causality and Statistical Learning*, 117 *Am. J. Sociology* 955 (2011).

For most questions of legal interest, a wide variety of models can be used. This is only to be expected, because the science does not dictate specific equations and causal relationships. In a strongly contested case, each side will have its own model (series of models), presented by its own expert. The experts often reach opposite conclusions. The dialogue might continue with an exchange about which models produce admissible results, or more convincing ones. Although assumptions about the form of the model or the error component are challenged in court from time to time, arguments more commonly revolve around the choice of variables. One model may be questioned because it omits variables that arguably should be included—for example, skill levels or prior evaluations in an employment discrimination case.¹⁹⁴ Another model may be challenged because it includes tainted variables reflecting past discriminatory behavior by the firm.¹⁹⁵ One expert may emphasize how remarkably well the fitted regression curve or surface fits; another expert may reply that the model, especially a complicated one, is ad hoc—it is tailored to the data in question and cannot be relied on to work well with other data arising from the same data-generating process.¹⁹⁶ The court or jury must decide which model—if any—fits the occasion.¹⁹⁷

The frequency with which regression models are used is no guarantee that they are the best choice for any particular problem.¹⁹⁸ Indeed, from one perspective, a regression or other statistical model may seem to be a marvel of mathematical rigor. From another perspective, the model could be a set of assumptions,

194. *E.g.*, *Bazemore v. Friday*, 478 U.S. 385 (1986); *In re Linerboard Antitrust Litig.*, 497 F. Supp. 2d 666 (E.D. Pa. 2007).

195. *E.g.*, *McLaurin v. Nat'l R.R. Passenger Corp.*, 311 F. Supp. 2d 61, 65–66 (D.D.C. 2004) (holding that the inclusion of two allegedly tainted variables was reasonable in light of an earlier consent decree); *see also In re Urethane Antitrust Litig.*, 166 F. Supp. 3d 501, 506 (D.N.J. 2016) (“courts have declined to fault a regression model for excluding variables that could have been manipulated by a defendant in furtherance of an antitrust conspiracy”).

196. *E.g.*, *Students for Fair Admissions, Inc. v. Univ. of N.C.*, 567 F. Supp. 3d 580, 622 (M.D.N.C. 2021) (“both experts agree that the degree of overfit is a factor they must take into account, the parties disagree on how to measure it”), *rev'd*, 600 U.S. 181 (2023); *In re Urethane Antitrust Litig.*, 166 F. Supp. 3d at 505 (predictions were admissible despite the argument “that the models in this case were able to accurately match prices during the benchmark period not because they are reliable, but because they are ‘overfit’”).

197. *E.g.*, *Chang v. Univ. of R.I.*, 606 F. Supp. 1161, 1207 (D.R.I. 1985) (“it is plain to the court that [defendant’s] model comprises a better, more useful, more reliable tool than [plaintiff’s] counterpart”); *Presseisen v. Swarthmore Coll.*, 442 F. Supp. 593, 619 (E.D. Pa. 1977) (“[E]ach side has done a superior job in challenging the other’s regression analysis, but only a mediocre job in supporting their own . . . and the Court is . . . left with nothing.”), *aff’d*, 582 F.2d 1275 (3d Cir. 1978).

198. *See, e.g.*, *In re Urethane Antitrust Litig.*, 166 F. Supp. 3d at 504–05 (although “there are an abundance of judicial decisions supporting the premise that regression models can be a reliable tool for measuring damages in an antitrust case,” plaintiff’s “regression models must stand on their own two feet”).

supported only by the say-so of the testifying expert. Intermediate judgments are also possible.¹⁹⁹

Data Science and Statistical Machine Learning

“Statistical science” has been described as “the discipline of learning about the world from data,”²⁰⁰ and “statistics” as the art of “learning from data.”²⁰¹ Yet, there are now journals,²⁰² university degree programs,²⁰³ and job positions²⁰⁴ dedicated—not to statistics—but to “data science.” Spurred by the availability of large and heterogeneous data collections, computer-intensive algorithmic procedures for making predictions and classifications are increasingly applied in science, business, and law enforcement.²⁰⁵ These procedures have often been termed “analytics,” “artificial intelligence,” and “machine learning.” How might testimony and reports from “data scientists” differ from more traditional data analyses of statisticians? What statistical considerations or methods should be used to validate the output from machine learning approaches? To supply background information for approaching these questions, this section describes the

199. See, e.g., David W. Peterson, *Reference Guide on Multiple Regression*, 36 *Jurimetrics J.* 213, 214–15 (1996) (review essay); see *supra* note 25 for references to a range of academic opinion. More recently, some investigators have turned to graphical models. However, these models have serious weaknesses of their own. See, e.g., David A. Freedman, *Graphical Models for Causation, and the Identification Problem*, 28 *Evaluation Rev.* 267 (2004), <https://doi.org/10.1177/0193841X04266432>.

200. Spiegelhalter, *supra* note 95, at 404 (Basic Books edition).

201. This is the subtitle of the Penguin Books edition of Spiegelhalter, *supra* note 95.

202. The *Harvard Data Science Review* (<https://perma.cc/T87V-VEP3>), for example, “aim[s] to publish content that help define and shape data science as a scientifically rigorous and globally impactful multidisciplinary field based on the principled and purposed production, processing, parsing, and analysis of data.”

203. U.S. News & World Rep., *Best Undergraduate Data Science Programs*, <https://www.usnews.com/best-colleges/rankings/computer-science/data-analytics-science>.

204. The Bureau of Labor Statistics has an occupational category for “data scientists” that excludes statisticians. U.S. Bureau of Labor Statistics, *Occupational Employment and Wages*, May 2021, <https://www.bls.gov/oes/current/oes152051.htm>.

205. Applications of particular legal interest include predicting criminal behavior at the individual level (see, e.g., Richard A. Berk & J. Bleich, *Statistical Procedures for Forecasting Criminal Behavior: A Comparative Assessment*, 12 *Criminology & Public Policy* 513 (2013), <https://doi.org/10.1111/1745-9133.12407>, classifying trace evidence such as toolmarks according to their source (see, e.g., ANSI/ASB 062, *Standard for Topography Comparison Software for Toolmark Analysis* (2021)), and forensic speaker recognition (see, e.g., Geoffrey Stewart Morrison & William C. Thompson, *Assessing the Admissibility of a New Generation of Forensic Voice Comparison Testimony*, 18 *Colum. Sci. & Tech. L. Rev.* 326, 345–46 (2017); Zhongxin Bai & Xiao-Lei Zhang, *Speaker Recognition Based on Deep Learning: An Overview*, 140 *Neural Networks* 65 (2021), <https://doi.org/10.1016/j.neunet.2021.03.004>).

general nature of “data science” and sketches the broad outlines of the statistical thinking that is crucial to evaluating results from machine learning.²⁰⁶

What Is Data Science?

“There is a wide variety of definitions and criteria for what constitutes data science,” making the term something of “a buzzword.”²⁰⁷ Typical definitions, such as “a collection of techniques used to extract value from data [that] rely on finding useful patterns, connections, and relationships within data,”²⁰⁸ do not distinguish data science from applied statistics.²⁰⁹ For centuries, statisticians have been concerned with discerning patterns and regularities in natural and social processes and with predicting uncertain outcomes.

Nonetheless, data science has emerged as a *combination* of statistics, computing, and informatics ideas about the collection, storage, analysis, visualization, and reporting of data. Advances in technology, computing, and data-storage capacity have produced a data explosion in commerce (partly driven by online shopping), science (massive sky surveys, environmental and climate data, genomic information, and much more), and digitalized voice, text, and pictures collected to support speech and image recognition.²¹⁰ The “data science” that “extracts value” from these new troves of data is a conglomeration of mathematically grounded modeling methods from traditional statistics and highly pragmatic procedures for making predictions and classifications. The latter methods often fall under the rubric of “machine learning” and are sketched in the next section.

What Is Machine Learning?

“Machine learning” refers to methods for using data to classify objects or patterns into categories, and to produce predictions of future outcomes or events. The

206. For a general discussion of AI and related societal implications, see James E. Baker & Laurie N. Hobert, *Reference Guide on Artificial Intelligence*, in this manual.

207. Vijay Kotu & Bala Deshpande, *Data Science: Concepts and Practice* 1 (2d ed. 2019).

208. *Id.*; see also Bradley Efron & Trevor Hastie, *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science* 451 (2016) (noting that “[t]he data science association defines a practitioner as one who ‘uses scientific methods to liberate and create meaning from raw data’”).

209. See Matthew A. Jay & Mario Cortina Borja, *Quomodo Dicitur “Data Science” Latine*, *Significance*, Aug. 2020, at 26.

210. Initially the increased amount of data was heralded as the dawn of a Big Data era involving Big Data specialists. See Francis X. Diebold, *What’s the Big Idea? “Big Data” and Its Origins*, *Significance*, Feb. 2021, at 36–37. “Big” can refer to the number of units (people, social media posts, galaxies, etc.) in a dataset (“examples” in a database). It also can refer to the number of variables for each unit (the “features” of the examples). Genomes are big—they have billions of base pairs—even if they come from only a handful of people.

equations or procedures rely on computers. Of course, classification and prediction are tasks that humans do, either intuitively or with experience and expert training. Thus, machine learning (ML) is a subfield of “artificial intelligence” (AI), which seeks to build machines that perform tasks we associate with intelligent agents.²¹¹

ML is closely related to statistics. Some methods that are packaged as ML or AI are nothing fundamentally new.²¹² For example, regression models can play a role. An easily understood example—predicting the first-year grade-point average (GPA) of a student entering law school from only one variable—illustrates some of the terminology. The school knows the LSAT score and the subsequent first-year GPA of every student in last year’s class. It can make predictions by fitting a straight line to those data points. As explained in the sections titled “Correlation and Regression” and “What Inferences Can Be Drawn from the Data?” above, the fitted line could have the mathematical form $GPA = \hat{a} + \hat{b} \times LSAT$, where $\hat{a}$ and $\hat{b}$ are the line’s intercept and slope as estimated from the data.²¹³ The estimates could come from equations for the values that minimize the sum of the squared differences between each data point and the fitted line. (That criterion, in turn, can be justified by a regression model in which $GPA = a + b \times LSAT + \epsilon$, where a and b are the unknown parameters and ϵ is a normally distributed error term.) With the fitted equation in place (and under the assumption that the relationship between the variables for last year’s students will continue to apply), the prediction of an applicant’s

211. AI is extremely broad in its scope and tools, with no single, commonly accepted definition. See, e.g., Stuart J. Russell & Peter Norvig, *Artificial Intelligence: A Modern Approach* (3d ed. 2010). Indeed, “so far at least, all the candidates [for a definition] have substantial flaws.” Richard A. Berk, *Artificial Intelligence, Predictive Policing, and Risk Assessment for Law Enforcement*, 4 Ann. Rev. Criminology 209, 211 (2021). For a time, AI focused on building symbolic representations of the world and developing rule- or case-based systems that could reason based on those representations. See, e.g., Philip Leith, *The Rise and Fall of the Legal Expert System*, 1 European J. L. & Tech. No. 1 (2010), <https://perma.cc/E8MF-JTFU>. Although statisticians and machine-learning researchers do not think of their work as necessarily being motivated by traditional AI goals such as elucidating how the brain processes information, the statistical and algorithmic methods have produced automated systems that work well for narrow tasks. These accomplishments are now designated “narrow” or “weak” AI. However, “the artificial intelligence of today is computer code written to accomplish a specific empirical task. Restated, it is just a computational procedure. In law enforcement settings, these procedures are by and large enhancements of activities that humans are already performing. Arguably, computers can do them faster and more accurately, but the internal operations are far removed from the way humans actually think. It is not even clear that the term intelligence applies.” Berk, *supra*, at 212.

212. Susan Athey & Guido W. Imbens, *Machine Learning Methods that Economists Should Know About*, 11 Ann. Rev. Econ. 685, 689 (2019), <https://doi.org/10.1146/annrev-economics-080217-053433> (observing that “[o]ne source of confusion is the use of new terminology in ML for concepts that have well-established labels in the older literatures” and providing examples in the context of regression analysis); Berk, *supra* note 211, at 223–24 (some crime-forecasting procedures presented as ML “are a rebranding of statistical tools developed more than 50 years ago”).

213. Additional predictors and other functional forms could be explored with the statistical procedures mentioned in the *Reference Guide on Multiple Regression and Advanced Statistical Models*.

score is simple arithmetic—one multiplication and one addition. The calculations, including the estimates of a and b , could be done by hand for a single law school, or the LSAT–GPA data could be entered into a spreadsheet or a statistical package of software that would generate the prediction equation and any desired prediction based on it. Next year, the machine will have new “training” data (another set of actual GPAs and LSATs) from which it can “learn” to make predictions (from a regression equation with updated, estimated values for the parameters a and b).

Prediction by simple linear regression is so well established it is unlikely to be packaged as machine learning. That phrase is more likely to be encountered with procedures that depart from such explicit modeling and that are said to be “algorithmic.”²¹⁴ The word itself is not very revealing. Algorithms are step-by-step instructions for computing something. They can be written down in English or another natural language, or depicted in a flow chart. Then they can be implemented on a computer after they are converted into a specific computer programming language (“code”).

The regression model-based procedure for predicting GPA entails two parts that can be called algorithmic. First, there was an algorithm for finding the particular regression line.²¹⁵ Second, this line enables us to write down an algorithm for calculating the predicted value. This second algorithm might go as follows: (1) input the applicant’s LSAT score; (2) multiply it by the value $\hat{b}$ found from all the data; add the value found for $\hat{a}$ to the result; (3) output the result of step (2) for the predicted GPA.

Because algorithms underlie all computation, it is not the existence of algorithms that makes ML distinctive. It is the willingness to develop algorithms for predictions or classifications without necessarily positing a statistical model of how the data arise. Statisticians historically relied on statistical models with a relatively small number of interpretable parameters (like the intercept a and the slope b of the simple regression line) and used manageably small datasets to estimate the values of these parameters. But models with many parameters are characteristic of ML.²¹⁶

To convey the rough idea, let’s imagine that instead of imposing the linear regression model on the LSAT–GPA data, we used the following algorithm: (1) compute the average GPA for the students with LSATs ranging from 120–130,

214. *E.g.*, Berk, *supra* note 211, at 12 (“An essential feature of artificial narrow intelligence is that it depends on algorithms, not models.”); Moritz Hardt & Benjamin Recht, *Patterns, Predictions, and Actions: Foundations of Machine Learning* 11 (2020) (“Machine learning is to a large extent the study of algorithmic prediction.”); Efron & Hastie, *supra* note 208, at 451 (“the emphasis is on the algorithmic processing of large data sets for the extraction of useful information, with the prediction algorithms as exemplars”).

215. The equations for estimating a and b in the regression model are solved by a particular series of additions, subtractions, and multiplications that are guaranteed to give the specific values $\hat{a}$ and $\hat{b}$ that minimize the sum of the squared deviations between each data point and the overall fitted line.

216. For discussion of how such procedures relate to more traditional statistical methods, see *Special Issue: Commentaries on Breimen’s Two Cultures Paper*, 7 *Observational Stud.* 1–234 (2021).

130–140, and so on through 170–180; (2) plot these averages as heights at the mid-points of the intervals (125, 135, . . . , 175); (3) connect them to form a jagged line, and (4) use this jagged line to make predictions. Of course, one might ask why widths of 10 LSAT points should be used. Why not use shorter or longer segments for computing the means? An optimization procedure that reserves some of the data to try out different possibilities could help identify which width to use.²¹⁷ The approach resembles engineering by trial and error.²¹⁸

Our jagged line would not be used in practice. There are better procedures for fitting a curve to a cloud of data points. The slice-compute-and-connect algorithm is here only to illustrate the *idea* of a highly data-driven procedure that leads to an algorithm (corresponding to the resulting jagged line) for making the predictions. The jagged line would hug the data more tightly than the single regression line with only two parameters. Maybe it would produce better predictions.

A number of important ML procedures do not yield any explicit equation (like the single regression line or the more complex jagged line) for making predictions or classifications. The two stages are done together, and the relevant variables are not necessarily pre-defined by the developer of the prediction model. Also, the features of the data that have the most influence on the output often are unknown.

What Statistical Questions Arise with Machine Learning Studies?

Although machine-learning advocates sometimes advertise that their methods make very few assumptions and provide great flexibility, they are still subject to the statistical and logical issues discussed in previous sections. Data quality,

217. We could reserve a random 10% of the data and develop the jagged line on the remaining 90% using a width of, say, two LSAT points. Then we check how well the fitted line works on the reserved 10% by finding the correlation between the fitted line's predictions and the reserved data points. We do this repeatedly, say 10 times, to get an average correlation for the jagged line with the two-point width. Then we return to the full training set and repeat this 10-fold train-test process for other values of the width. Finally, we pick the width with the highest correlation ("cross-validity") and apply it to the full training set (100% of the data). That gives the prediction line.

218. Moritz Hardt & Benjamin Recht, *Patterns, Predictions, and Actions: Foundations of Machine Learning* 146 (2020) ("Methodologically, much of modern machine learning practice rests on a variant of *trial and error*, which we call the *train-test paradigm*. Practitioners repeatedly build models using any number of heuristics and test their performance to see what works. Anything goes as far as training is concerned, subject only to computational constraints, so long as the performance looks good in testing.").

robustness, and transparency are a few of the issues that can arise.²¹⁹ We consider these three in turn.²²⁰

Is the Dataset Appropriate and of Sufficient Quality?

To begin with, as with traditional statistical methods, care is needed in the assembly of the data used to build and test the algorithms. Algorithms must be developed on data that are accurate and representative of the type of data to which they will be applied. For example, a study on diagnosing autism using ML excluded the data corresponding to borderline cases of autism. That made the test set unrepresentative of the general population; moreover, when a dataset that included the difficult-to-diagnose cases was used, the accuracy declined.²²¹

Is the Predictor or Classifier Robust?

Second, robustness is also a vital issue. The goal of ML is not merely to find some complex pattern in the data; it is to discover *generalizable* patterns. For instance, a complex equation using many characteristics of each student to predict law school GPA could fit last year's data splendidly, but it might be idiosyncratic to the "training" data. In other words, it might not generalize to next year's data. This is known as overfitting. It can occur with traditional statistical techniques but is a greater risk with the high-dimensional (many variables) algorithmic approaches

219. For reviews of ML studies identifying frequent mistakes or departures from good practice, see Sayash Kapoor & Arvind Narayanan, *Leakage and the Reproducibility Crisis in Machine-learning-based Science*, 4 *Patterns* 100804 (2023), <https://doi.org/10.1016/j.patter.2023.100804>; Michael Roberts et al., *Common Pitfalls and Recommendations for Using Machine Learning to Detect and Prognosticate for COVID-19 Using Chest Radiographs and CT Scans*, 3 *Nature Machine Intelligence* 199 (2021); Laure Wynants et al., *Prediction Models for Diagnosis and Prognosis of Covid-19: Systematic Review and Critical Appraisal*, 369 *Brit. Med. J.* m1328 (2020), <https://doi.org/10.1136/bmj.m1328>. The extent to which published studies using ML-methods produce results that are superior to more traditional statistical methods is an active area of research. See, e.g., Mohammad Ziaul Islam Chowdhury et al., *Prediction of Hypertension Using Traditional Regression and Machine Learning Models: A Systematic Review and Meta-Analysis*, 17 *PLoS ONE* e0266334 (2022), <https://doi.org/10.1371/journal.pone.0266334>; Xuan Song et al., *Comparison of Machine Learning and Logistic Regression Models in Predicting Acute Kidney Injury: A Systematic Review and Meta-Analysis*, 151 *Int'l J. Med. Informatics* 104484 (2021), <https://doi.org/10.1016/j.ijmedinf.2021.104484>.

220. The choice of statistics to indicate the accuracy of a predictor or classifier is another important topic.

221. Daniel Bone et al., *Applying Machine Learning to Facilitate Autism Diagnostics: Pitfalls and Promises*, 45 *J. Autism & Developmental Disorders* 1121 (2015), <https://doi.org/10.1007/s10803-014-2268-6>.

that characterize modern machine learning. “Internal validation” can be achieved by holding out some data in the training set as the algorithm is developed and measuring how well the algorithm predicts or classifies those test data.²²² But such internal validation is not sufficient. “External validation” requires testing the algorithm on data from a different source, and results from those tests should be reported. Furthermore, even if external validation has been performed at one point in time, confidence in continued predictive accuracy requires ongoing efforts to ensure that the relationships among variables in the data are not changing over time (or to revise the equations or algorithms if they are).

Is the Predictor or Classifier Too Opaque?

Finally, ML algorithms can be difficult to interpret, especially when the patterns that are “learned” may not be represented in easily decoded forms. This can make it hard to tell when the data environment is changing in ways that will lead to poorer performance. It also can mask the fact that the algorithm has access to features that should not be part of the data. For example, when researchers rushed to apply ML methods to COVID-19 diagnosis from chest X-rays and CT scans, many of them unwittingly used a dataset that contained chest scans of children who did not have COVID as their examples of what non-COVID cases looked like. As a result, the ML diagnoses were relying on features identifying children rather than COVID.²²³ Concerns such as these are prompting the development of methods for gaining insight into how opaque ML systems are functioning²²⁴ and to lists of criteria for judging when ML algorithms are most likely to be trustworthy.²²⁵

222. This is a very rough description of cross-validation with data available when developing the algorithm. For discussion of data-splitting and resampling techniques, see, for example, Max Kuhn & Kjell Johnson, *Applied Predictive Modeling* 61–92 (2013).

223. *Id.*

224. Berk, *supra* note 211, at 231.

225. E.g., Kapoor & Narayanan, *supra* note 219; Russell A. Poldrack et al., *Establishment of Best Practices for Evidence for Prediction: A Review*, 77 JAMA Psychiatry 534 (2020), <https://doi.org/10.1001/jamapsychiatry.2019.3671>; Roberts et al., *supra* note 219.

Appendix: Conditional Probability and Bayes' Rule

What Do Probabilities Apply To?

The mathematical theory of probability consists of theorems derived from axioms and definitions. Mathematical reasoning is seldom controversial, but there may be disagreement as to how the theory should be applied. For example, statisticians may differ on the interpretation of data in specific applications. Moreover, there are two main schools of thought about the foundations of statistics: frequentist and Bayesian (also called objectivist and subjectivist).²²⁶

Frequentists see probabilities as empirical facts. When a fair coin is tossed, the probability of heads is taken to be $1/2$; this value is justified by the argument that if the experiment is repeated a large number of times, the coin will land heads about one-half the time. If a fair die is rolled, the probability of getting an ace (one spot) is $1/6$. If the die is rolled many times, an ace will turn up about one-sixth of the time.²²⁷ Generally, if a chance experiment can be repeated, the relative frequency of an event approaches (in the long run) its probability. By contrast, a Bayesian considers probabilities as representing not facts but degrees of belief: in whole or in part, probabilities are subjective.²²⁸

226. The extent to which statisticians using Bayesian methods are overtly subjective in their analyses varies. "Objective Bayesians" use Bayes' rule without eliciting prior probabilities from subjective beliefs. One strategy is to use preliminary data to estimate the prior probabilities and then apply Bayes' rule to that empirical distribution. This "empirical Bayes" procedure avoids the charge of subjectivism at the cost of departing from a fully Bayesian framework. With ample data, it can be effective, and the estimates or inferences can be understood in frequentist terms. Another "objective" approach is to use "noninformative" priors that are supposed to be independent of all data and prior beliefs. However, the choice of such priors can be questioned, and the approach has been contested by frequentists and subjective Bayesians. *E.g.*, Joseph B. Kadane, *Is "Objective Bayesian Analysis" Objective, Bayesian, or Wise?*, 1 *Bayesian Analysis* 433 (2006), <https://perma.cc/CSN9-35G5>; Jon Williamson, *Philosophies of Probability*, in *Philosophy of Mathematics* 493 (Andrew Irvine ed., 2009) (discussing the challenges to objective Bayesianism).

227. Probabilities may be estimated from relative frequencies, but probability itself is a subtler idea. For example, suppose a computer prints out a sequence of ten letters H and T (for heads and tails), which alternate between the two possibilities H and T as follows: H T H T H T H T H T. The relative frequency of heads is $5/10$ or 50%, but it is not at all obvious that the chance of an H at the next position is 50%. There are difficulties in both the subjectivist and objectivist positions. See Freedman, *supra* note 167.

228. "Subjective" is not necessarily the same as arbitrary. The degrees of belief must obey the same mathematical rules (axioms and theorems) as relative frequencies do, and the personal choices for particular numbers can be subjected to interpersonal standards for acceptance by other individuals.

What Are Conditional Probabilities?

Conditional probability is the probability of one event given that another has occurred. For example, suppose a fair coin is tossed twice. One event is that the coin will land heads up both times (HH). Another event is that at least one H will be seen. Before the coin is tossed, there are four possible, equally likely, outcomes: HH, HT, TH, TT. So the probability of HH is 1/4. However, if we know that at least one head has been obtained, then we can rule out two tails TT. In other words, given that at least one H has been obtained, the conditional probability of TT is 0, and the first three outcomes have conditional probability 1/3 each. In particular, the conditional probability of HH is 1/3. This is usually written as $P(\text{HH}|\text{at least one H}) = 1/3$. More generally, the probability of an event C is denoted $P(C)$; the conditional probability of D given C is written as $P(D|C)$.

Two events C and D are independent if the conditional probability of D given that C occurs is equal to the conditional probability of D given that C does not occur. Using $\sim C$ to denote the event that C does not occur, C and D are independent if $P(D|C) = P(D|\sim C)$. If C and D are independent, then the probability that both occur is equal to the product of the probabilities:

$$P(C \text{ and } D) = P(C) \times P(D) \quad (3)$$

This is the multiplication rule (or product rule) for independent events. If events are dependent, then conditional probabilities must be used:

$$P(C \text{ and } D) = P(C) \times P(D|C) \quad (4)$$

This is the multiplication rule for dependent events. Statisticians caution against, but sometimes succumb to, using the multiplication rule for independent events when the events are dependent.²²⁹

What Is Bayes' Rule?

If we use probabilities to describe uncertainty regarding hypotheses as well as to describe events, we can use the formulas for conditional probabilities to obtain an equation known as Bayes' rule that has been proposed to guide legal factfinding. If

229. E.g., William Kruskal, *Miracles and Statistics: The Casual Assumption of Independence*, 83 J. Am. Stat. Ass'n 929 (1988), <https://doi.org/10.2307/2290117>. A famous case of multiplication without apparent attention to dependence is *People v. Collins*, 438 P.2d 33 (Cal. 1968). For more examples, see *infra* note 238.

two mutually exclusive hypotheses H_0 and H_1 describe all the ways that an event A could occur,²³⁰ it is easy to show that²³¹

$$P(H_1 | A) = \frac{P(A | H_1)P(H_1)}{P(A | H_0)P(H_0) + P(A | H_1)P(H_1)}. \quad (5)$$

This is one way of expressing Bayes' rule. It yields the conditional probability of hypothesis H_1 given that event A has occurred.

For a stylized example in a criminal case, suppose that blood from a single source is found at the scene of a crime and that the defendant in the case has type A blood. It is natural to have one hypothesis H_0 propose that the blood came from a person other than the defendant; the other hypothesis H_1 is that the blood came from the defendant; the event A is the event that blood from the crime scene is found to be type A. Then $P(H_0)$ is the “prior probability” of H_0 , based on subjective judgment of the other evidence and background information in the case, while $P(H_0|A)$ is the “posterior probability”—updated from the prior probability using the data on the blood.

Type A blood occurs in 42% of the population. Assuming that the laboratory's findings are always correct, $P(A|H_0) = 0.42$.²³² Because the defendant has type A blood, if he is the source, event A will occur: $P(A|H_1) = 1$. Suppose the

230. “Mutually exclusive” means that either one hypothesis or the other—but not both—is true.

231. We use the multiplication rule (4) to find that

$$P(A \text{ and } H_0) = P(A|H_0) P(H_0) \quad (6)$$

and

$$P(A \text{ and } H_1) = P(A|H_1) P(H_1). \quad (7)$$

Moreover, if A can occur only when either H_0 or H_1 is true, then

$$P(A) = P(A \text{ and } H_0) + P(A \text{ and } H_1). \quad (8)$$

The multiplication rule (4) also shows that

$$P(H_1|A) = P(A \text{ and } H_1) / P(A). \quad (9)$$

Using (7) to evaluate $P(A \text{ and } H_1)$ in the numerator of (9), and (6), (7), and (8) to evaluate $P(A)$ in the denominator gives equation (5) in the text.

232. Not all statisticians would accept the identification of a population frequency with $P(A|H_0)$. Indeed, H_0 has been translated into a hypothesis that the true donor has been selected from the population at random (i.e., in a manner that is uncorrelated with blood type). This step needs justification.

prior probabilities are $P(H_0) = P(H_1) = 0.5$. According to (5), the posterior probability that the blood is from the defendant is

$$P(H_1 | A) = \frac{1 \times 0.5}{(0.42 \times 0.5) + (1 \times 0.5)} = 0.70 \quad (10)$$

Thus, the data increase the probability that the blood is the defendant's. The probability went up from the prior value of $P(H_1) = 0.50$ to the posterior value of $P(H_1 | A) = 0.70$.

Equation (5) can be rewritten as follows:

$$\frac{P(H_1 | A)}{P(H_0 | A)} = \frac{P(A | H_1)}{P(A | H_0)} \times \frac{P(H_1)}{P(H_0)}. \quad (11)$$

This form gives us a convenient way to express the idea that the data change the chances in favor of H_1 as opposed to H_0 . The ratio of the two prior probabilities is the odds in favor of H_1 as opposed to H_0 .²³³ In terms of odds, the version of Bayes' rule in (5) becomes²³⁴

$$\text{Posterior odds} = \text{Likelihood ratio} \times \text{Prior odds} \quad (12)$$

where the "likelihood ratio" is the probability of the data (the crime-scene blood is type A) under the hypotheses H_1 divided by the probability of the same data under the other hypothesis H_0 . This ratio states how many times more (or less) probable the data are under H_1 as opposed to H_0 . Here, it is $1/0.42 = 2.31$. We would expect the data (type A blood) more than twice as often for the defendant than for a randomly selected individual; to that extent, it supports the hypothesis that the defendant is the source. It is more compatible with that hypothesis than the alternative being considered.

Used with Bayes' rule, this likelihood ratio means that the data increase the prior odds by a factor of a little more than 2, from 1:1 to about 2.31:1. Odds of 2.31 to 1 are the same as a probability of $2.31/(2.31 + 1) = 0.70$. Equations (5), (11), and (12) are all forms of the same fundamental relationship. In the context of Bayes' rule, the likelihood ratio is called a Bayes' factor.²³⁵ It is the *change* in the

233. If the odds in favor of an event or a hypothesis are j to k , the probability is j divided by $j + k$. For example, favorable odds of seven to three (more compactly written as 7:3 or 7/3) indicate a probability of $7/10 = 0.70$.

234. Unlike equation (5), equations (11) and (12) hold even when other hypotheses could account for A .

235. See Kaye, *supra* note 172; cf. Stern, *supra* note 110 (summarizing frequentist, likelihood, and Bayesian frameworks for statistical inference). The likelihood ratio for binary test results or decisions, such as those in Table 1, is the true positive proportion (sensitivity) divided by the false positive probability ($1 - \text{specificity}$). Low sensitivity degrades the value of a positive test result,

odds, which can be quite different from—and should not be confused with—the posterior odds.²³⁶

The same ideas can be applied more broadly to the various statistical models that have been illustrated in this reference guide. There are Bayesian approaches to fitting regression or other statistical models to make predictions and to obtain estimates of population parameters.²³⁷ But whatever mode of statistical inference is employed, assessing probabilities, conditional probabilities, and independence is not entirely straightforward. Inquiry into the basis for expert judgment may be useful, and casual assumptions about independence should be questioned.²³⁸

which is why the false-positive probability is not a complete measure of the probative value of a positive finding.

236. For cases in which this mistake has been made, see David H. Kaye, *Reference Guide on Human DNA Identification Evidence*, “Presentation of LR’s” section, in this manual. The numerical disparity between the posterior odds and the Bayes’ factor will depend on its magnitude of the Bayes’ factor and on the prior odds. For example, the two are numerically equal when the prior odds are 1:1 (a prior probability of 50%). But if we insert prior odds of 1:10 into (12), the posterior odds become 2.31:10, or about 7:30. For these lower prior odds, the posterior probability is about $7/37 = 0.19$. Even though the blood-type evidence supports the same-source hypothesis over the different-source hypothesis by a factor of more than two, the same-source hypothesis remains improbable.

237. Modern treatments of Bayesian methods include Donald A. Berry, *Statistics: A Bayesian Perspective* (1995); Andrew Gelman et al., *Bayesian Data Analysis* (3d ed. 2013); Jeff Gill, *Bayesian Methods: A Social and Behavioral Sciences Approach* (3d ed. 2015); John K. Kruschke, *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan* (2d ed. 2015); Richard McElreath, *Statistical Rethinking: A Bayesian Course with Examples in R and STAN* (2d ed. 2020).

238. For problematic assumptions of independence in litigation, see, for example, *Wilson v. State*, 803 A.2d 1034 (Md. 2002) (error to admit multiplied probabilities in a case involving two deaths of infants in same family); 1 McCormick, *supra* note 1, § 210; see also *supra* note 36 (on census litigation).

Glossary of Terms

The following definitions are adapted from a variety of sources, including Michael O. Finkelstein & Bruce Levin, *Statistics for Lawyers* (2d ed. 2001), David A. Freedman et al., *Statistics* (4th ed. 2007), and David Spiegelhalter, *The Art of Statistics: How to Learn from Data* (2019).

absolute value. Size, neglecting sign. The absolute value of +2.7 is 2.7; so is the absolute value of −2.7.

adjust for. See control for.

alpha (α). A symbol often used to denote the probability of a Type I error. See Type I error; size. Compare beta.

alternative hypothesis. A statistical hypothesis that is contrasted with the null hypothesis in a significance test. See statistical hypothesis; significance test.

area sample. A probability sample in which the sampling frame is a list of geographical areas. That is, the researchers make a list of areas, choose some at random, and interview people in the selected areas. This is a cost-effective way to draw a sample of people. See probability sample; sampling frame.

arithmetic mean. See mean.

average. See mean. “Average” often is used generically, to refer to a single representative or central value, such as the mean, median, or mode for a set of numbers.

Bayes factor. A ratio that indicates how much the data changes the prior odds. See Bayes’ rule.

Bayes’ rule. In its simplest form, an equation involving conditional probabilities that relates a “prior probability” known or estimated before collecting certain data to a “posterior probability” that reflects the impact of the data on the prior probability.

Bayesian inference. In Bayesian statistical inference, “the prior” expresses degrees of belief about various hypotheses or parameters before data are collected. Data are then collected according to some statistical model; at least, the model represents the investigator’s beliefs. Bayes’ rule combines the prior probability with the data to yield the posterior probability, which expresses the investigator’s beliefs about the hypotheses or parameters, given the data. See Appendix. Compare frequentist.

beta (β). A symbol sometimes used to denote power, and sometimes to denote the probability of a Type II error. See Type II error; power. Compare alpha.

between-observer variability. Differences that occur when two or more observers measure the same thing. Compare within-observer variability.

bias. Also called systematic error. A systematic tendency for an estimate to be too high or too low. An estimate is unbiased if the bias is zero. (Bias does not mean prejudice, partiality, or discriminatory intent.) See nonsampling error. Compare sampling error.

bias-variance trade-off. Some estimators computed from a sample have less variability across samples than an unbiased estimator, making it reasonable to accept some bias in order to increase precision (reduce the standard error of the estimate). Similarly, when fitting a model for prediction, increasing complexity will eventually lead to a model that has less bias, in the sense that it has greater potential to adapt to details of the underlying process, but more variance, since there is not enough data to be confident about the parameters in the model. These elements need to be traded off to avoid over-fitting.

big data. A huge volume of data, often from a variety of sources such as images, social media accounts, or transactions, having a high velocity of acquisition and a possible lack of veracity due to its routine collection.

bin. A class interval in a histogram. See class interval; histogram.

binary variable. A variable that has only two possible values (e.g., sex). Called a dummy variable when the two possible values are 0 and 1.

binomial distribution. A distribution for the number of occurrences in repeated, independent “trials” where the probabilities are fixed. For example, the number of heads in 100 tosses of a coin follows a binomial distribution. When the probability is not too close to 0 or 1 and the number of trials is large, the binomial distribution has about the same shape as the normal distribution. See normal distribution; Poisson distribution.

blind. See double-blind experiment.

bootstrap. Also called resampling; Monte Carlo method. A procedure for estimating sampling error by constructing a simulated population on the basis of the sample, then repeatedly drawing samples from the simulated population.

categorical data; categorical variable. See qualitative variable. Compare quantitative variable.

central limit theorem. Mathematical theorem showing that under suitable conditions, the probability histogram for a sum (or average or rate) will follow the normal curve. See histogram; normal curve.

chance error. See random error; sampling error.

chi-squared (χ^2). The chi-squared statistic measures the distance between the data and expected values computed from a statistical model. If the chi-squared statistic is too large to explain by chance, the data contradict the model. The definition of “large” depends on the context. See statistical hypothesis; significance test.

class interval. Also, bin. The base of a rectangle in a histogram; the area of the rectangle shows the frequency or relative frequency of observations in the class interval. See histogram.

cluster sample. A type of random sample. For example, investigators might take households at random, then interview all people in the selected households. This is a cluster sample of people: A cluster consists of all the people in a selected household. Generally, clustering reduces the cost of interviewing. See multistage cluster sample.

coefficient of determination. A statistic (more commonly known as *R*-squared) that describes how well a regression equation fits the data. See *R*-squared.

coefficient of variation. A statistic that measures spread relative to the mean: SD/mean, or SE/expected value. See expected value; mean; standard deviation; standard error.

collinearity. See multicollinearity.

conditional probability. The probability that one event will occur given that another has occurred.

confidence coefficient. See confidence interval.

confidence interval. An estimate, expressed as a range, for a parameter. For estimates such as averages or rates computed from large samples, a 95% confidence interval is the range from about two standard errors below to two standard errors above the estimate. Intervals obtained this way cover the true value about 95% of the time, and 95% is the confidence level or the confidence coefficient. See central limit theorem; standard error.

confidence level. See confidence interval.

confounding variable; confounder. A confounder is correlated with the independent variable and the dependent variable. An association between the dependent and independent variables in an observational study may not be causal, but may instead be due to confounding. See controlled experiment; observational study.

consistent estimator. An estimator that tends to become more and more accurate as the sample size grows. Inconsistent estimators, which do not become more accurate as the sample gets larger, are frowned upon by statisticians.

content validity. The extent to which a skills test is appropriate to its intended purpose, as evidenced by a set of questions that adequately reflect the domain being tested. See validity. Compare reliability.

continuous variable. A variable that has arbitrarily fine gradations, such as a person's height. Compare discrete variable.

control for. Statisticians may control for the effects of confounding variables in nonexperimental data by making comparisons for smaller and more

homogeneous groups of subjects, or by entering the confounders as explanatory variables in a regression model. To “adjust for” is perhaps a better phrase in the regression context, because in an observational study the confounding factors are not under experimental control; statistical adjustments are an imperfect substitute. See regression model.

control group. See controlled experiment.

controlled experiment. An experiment in which the investigators determine which subjects are put into the treatment group and which are put into the control group. Subjects in the treatment group are exposed by the investigators to some influence (the treatment); those in the control group are not so exposed. For example, in an experiment to evaluate a new drug, subjects in the treatment group are given the drug, and subjects in the control group are given some other therapy; the outcomes in the two groups are compared to see whether the new drug works. Randomization—that is, randomly assigning subjects to each group—is usually the best way to ensure that any observed difference between the two groups comes from the treatment rather than from preexisting differences. In many situations, a randomized controlled experiment is impractical, and investigators must then rely on observational studies. Compare observational study.

convenience sample. A nonrandom sample of units, also called a grab sample. Such samples are easy to take but may suffer from serious bias. Typically, mall samples are convenience samples.

correlation coefficient. A number between -1 and 1 that indicates the extent of the linear association between two variables. Often, the correlation coefficient is abbreviated as r .

covariance. A quantity that describes the statistical interrelationship of two variables. Compare correlation coefficient; standard error; variance.

covariate. A variable that is related to other variables of primary interest in a study; a measured confounder; a statistical control in a regression equation.

credible interval. A range of values obtained through a Bayesian analysis that contains a parameter with a specified degree of belief.

criterion. The variable against which an examination or other selection procedure is validated. See validity.

data. Observations or measurements, usually of units in a sample taken from a larger population.

data science. The study and application of techniques for deriving insights from data, including constructing algorithms for prediction. Traditional statistical science forms part of data science, which also includes a strong element of computer science and data management.

degrees of freedom. See t -test.

dependence. Two events are dependent when the probability of one is affected by the occurrence or non-occurrence of the other. Compare independent; dependent variable.

dependent variable. Also called outcome variable. Compare independent variable.

descriptive statistics. Like the mean or standard deviation, used to summarize data.

differential validity. Differences in validity across different groups of subjects. See validity.

discrete variable. A variable that has only a small number of possible values, such as the number of automobiles owned by a household. Compare continuous variable.

distribution. See frequency distribution; probability distribution; sampling distribution.

disturbance term. A synonym for error term.

double-blind experiment. An experiment with human subjects in which neither the diagnosticians nor the subjects know who is in the treatment group or the control group. This is accomplished by giving a placebo treatment to patients in the control group. In a single-blind experiment, the patients do not know whether they are in treatment or control; the diagnosticians have this information.

dummy variable. See indicator variable.

econometrics. Statistical study of economic issues.

epidemiology. Statistical study of disease or injury in human populations.

error term. The part of a statistical model that describes random error or chance variation, i.e., the impact of chance factors unrelated to variables in the model. In econometrics, the error term is called a disturbance term.

estimator. A sample statistic used to estimate the value of a population parameter. For example, the sample average commonly is used to estimate the population average. The term “estimator” connotes a statistical procedure, whereas an “estimate” connotes a particular numerical result.

expected value. The expected value of a random variable is the weighted average of the possible values; the weights are the probabilities of the values. For example, the spots that turn up when a pair of dice are tossed is a random variable, and the expected value is

$$\begin{aligned} & \frac{1}{36} \times 2 + \frac{2}{36} \times 3 + \frac{3}{36} \times 4 + \frac{4}{36} \times 5 + \frac{5}{36} \times 6 + \frac{6}{36} \times 7 \\ & + \frac{5}{36} \times 8 + \frac{4}{36} \times 9 + \frac{3}{36} \times 10 + \frac{2}{36} \times 11 + \frac{1}{36} \times 12 \end{aligned}$$

which is equal to 7.

experiment. See controlled experiment; randomized controlled experiment.
Compare observational study.

explanatory variable. See independent variable; regression model.

external validity. See validity.

factors. See independent variable.

false negative. In a statistical analysis, the decision not to reject the null hypothesis when it is in fact false. In studies of forensic-science identification procedures, it is clearer to use terminology such as “false identification” or “false exclusion” to specify the type of incorrect decision being addressed.

false positive. In a statistical analysis, the decision to reject the null hypothesis when it is in fact true. In studies of forensic-science identification procedures, it is clearer to use terminology such as “false identification” or “false exclusion” to specify the type of incorrect decision being addressed.

Fisher’s exact test. A statistical test for comparing two sample proportions. For example, take the proportions of white and Black employees getting a promotion. An investigator may wish to test the null hypothesis that promotion does not depend on race. Fisher’s exact test is one way to arrive at a p -value. The calculation is based on the hypergeometric distribution. See hypergeometric distribution; p -value; significance test; statistical hypothesis.

fitted value. See residual.

fixed significance level. Also alpha; size. A preset level, such as 5% or 1%; if the p -value of a test falls below this level, the result is deemed statistically significant. See significance test. Compare observed significance level; p -value.

frequency; relative frequency. Frequency is the number of times that something occurs; relative frequency is the number of occurrences, relative to a total. For example, if a coin is tossed 1,000 times and lands heads 517 times, the frequency of heads is 517; the relative frequency is 0.517, or 51.7%.

frequency distribution. Shows how often specified values occur in a dataset.

frequentist. Describes statisticians who view probabilities as objective properties of a system that can be measured or estimated and who use statistical procedures justified by their performance in repeated samples from the population of interest. Compare Bayesian inference. See Appendix.

Gaussian distribution. A synonym for normal distribution.

general linear model. Expresses the dependent variable as a linear combination of the independent variables plus an error term whose components may be dependent and have differing variances. See error term; linear combination; variance. Compare regression model.

grab sample. See convenience sample.

heteroscedastic. See scatter diagram.

highly significant. See *p*-value; practical significance; significance test.

histogram. A plot showing how observed values fall within specified intervals, called bins or class intervals. Generally, matters are arranged so that the area under the histogram for each class interval gives the frequency or relative frequency of data in that interval. With a probability histogram, the area gives the chance of observing a value that falls in the interval.

homoscedastic. See scatter diagram.

hypergeometric distribution. Suppose a sample is drawn at random, without replacement, from a finite population. How many times will items of a certain type come into the sample? The hypergeometric distribution gives the probabilities. Compare Fisher's exact test.

hypothesis. See alternative hypothesis; null hypothesis; one-sided hypothesis; significance test; statistical hypothesis; two-sided hypothesis.

hypothesis test. Also, null hypothesis test; statistical test; test of significance. Involves formulating a statistical hypothesis (the null hypothesis) and an alternative hypothesis; devising a test statistic; and defining a critical region in which the test statistic prompts the rejection of the null hypothesis. The critical region (also called a rejection region) is constructed so that the probability of rejecting the null hypothesis—if the hypothesis is true—is no larger than some preestablished value (often $\alpha = 5\%$). Hypothesis tests often are performed by computing the *p*-value of the observed test statistic and comparing it to the preset value for rejection. See also significance test.

identically distributed. Random variables are identically distributed when they have the same probability distribution. For example, consider a box of numbered tickets. Draw tickets at random with replacement from the box. The draws will be independent and identically distributed.

independence. Also, statistical independence. Events are independent when the probability of one is unaffected by the occurrence or nonoccurrence of the other. Compare conditional probability; dependence; independent variable; dependent variable.

independent variable. Independent variables (also called explanatory variables, predictors, or risk factors) represent the causes and potential confounders in a statistical study of causation; the dependent variable represents the outcome or effect. In an observational study, independent variables may be used to divide the population into smaller and more homogenous groups ("stratification"). In a regression model, the independent variables are used to predict the dependent variable. For example, the unemployment rate has been used as the independent variable in a model for predicting the crime rate; the unemployment rate is the independent variable in this model, and the crime rate is the

dependent variable. The distinction between independent and dependent variables is unrelated to statistical independence. See regression model. Compare dependent variable; dependence; independence.

indicator variable. Variable created to assign numerical values to categorical or nominal data. The most common approach is to use a dummy variable that takes only the values 0 or 1 to distinguish one group of interest from another. See binary variable; regression model.

internal validity. See validity.

interquartile range. Difference between 25th and 75th percentile. See percentile.

interval estimate. A confidence interval, or an estimate coupled with a standard error. See confidence interval; standard error. Compare point estimate.

least squares. See least squares estimator; regression model.

least squares estimator. An estimator that is computed by minimizing the sum of the squared residuals. See residual.

level. The level of a significance test is denoted alpha (α). See alpha; fixed significance level; observed significance level; p -value; significance test.

likelihood. Colloquially, the chance that something will happen. In statistical models, the probability of observing a specific value of an observable quantity; may depend on the values of parameters.

likelihood ratio. A ratio of the probability of observing a specific value of an observable quantity under different hypotheses or parameter values.

linear combination. To obtain a linear combination of two variables, multiply the first variable by some constant, multiply the second variable by another constant, and add the two products. For example, $2u + 3v$ is a linear combination of u and v .

linear regression. A statistical model relating a continuous outcome or response to independent variable(s) or predictor(s). The dependent variable is modeled as a linear combination of the independent ones (or a transformed version of them) plus an error term whose values are randomly distributed in some fashion.

list sample. See systematic sample.

logistic regression. A statistical model relating a binary outcome or response to independent variable(s) or predictor(s). The logarithm of the odds of an event occurring is modelled as a linear combination of the independent variables plus a randomly distributed error term.

loss function. Statisticians may evaluate estimators according to a mathematical formula involving the errors—that is, differences between actual values and estimated values. The “loss” may be the total of the squared errors, or the total of the absolute errors, etc. Loss functions seldom quantify real losses,

but may be useful summary statistics and may prompt the construction of useful statistical procedures. Compare risk.

lurking variable. See confounding variable.

Markov chain. A random process describing a sequence of variables or events in which the probability of each variable or event depends only on the values of the immediately preceding variables or events.

mean. Also, the average; the expected value of a random variable. The mean gives a way to find the center of a batch of numbers: Add the numbers and divide by how many there are. Weights may be employed, as in “weighted mean” or “weighted average.” See random variable. Compare median; mode.

measurement validity. See validity. Compare reliability.

median. The median, like the mean, is a way to find the center of a batch of numbers. The median is the 50th percentile. Half the numbers are larger, and half are smaller. (To be very precise: at least half the numbers are greater than or equal to the median; at least half the numbers are less than or equal to the median; for small datasets, the median may not be uniquely defined.) Compare mean; mode; percentile.

meta-analysis. Attempts to combine information from all studies on a certain topic. For example, in the epidemiological context, a meta-analysis may attempt to provide a summary odds ratio and confidence interval for the effect of a certain exposure on a certain disease.

metrology. The scientific study of measurement.

mode. The most common number in a batch of numbers. Compare mean; median.

model. See probability model; regression model; statistical model.

Monte Carlo method. Relies on random sampling to estimate quantities of interest. The quantities of interest may be stochastic or deterministic. Monte Carlo methods give approximate solutions to many mathematical and statistical problems for which no simple analytical solution is available.

multicollinearity. Also, collinearity. The existence of correlations among the independent variables in a regression model. See independent variable; regression model.

multiple comparison. Making several statistical tests on the same dataset. Multiple comparisons complicate the interpretation of a p -value. For example, if twenty divisions of a company are examined, and one division is found to have a disparity significant at the 5% level, the result is not surprising; indeed, it would be expected under the null hypothesis. Compare p -value; significance test; statistical hypothesis.

multiple correlation coefficient. A number that indicates the extent to which one variable can be predicted as a linear combination of other variables. Its

magnitude is the square root of R -squared. See linear combination; R -squared; regression model. Compare correlation coefficient.

multiple regression. A regression equation that includes two or more independent variables. See regression model. Compare simple regression.

multistage cluster sample. A probability sample drawn in stages, usually after stratification; the last stage will involve drawing a cluster. See cluster sample; probability sample; stratified random sample.

multivariate methods. Methods for fitting models with multiple variables; in statistics, multiple response variables; in other fields, multiple explanatory variables. See regression model.

natural experiment. An observational study in which treatment and control groups have been formed by some natural development, but the assignment of subjects to groups is akin to randomization. See observational study. Compare controlled experiment.

nonparametric method. A method for testing a hypothesis or obtaining an interval that does not require knowledge of the form of the underlying parent population; also called a distribution-free method.

nonresponse bias. Systematic error created by differences between respondents and nonrespondents. If the nonresponse rate is high, this bias may be severe.

nonsampling error. A catch-all term for sources of error in a survey, other than sampling error. Nonsampling errors cause bias. One example is selection bias: The sample is drawn in a way that tends to exclude certain subgroups in the population. A second example is nonresponse bias: People who do not respond to a survey are usually different from respondents. A final example: Response bias arises if the interviewer uses a loaded question (or for other reasons).

normal distribution. Also, Gaussian distribution. Describes the probability density of certain continuous variables. The family of normal curves has two parameters: the mean and the standard deviation. The equation for the normal probability density function is given in the section “The normal curve and large samples.” Terminology notwithstanding, there need be nothing wrong with a distribution that differs from normal.

null hypothesis. For example, a hypothesis that there is no difference between two groups from which samples are drawn. See significance test; statistical hypothesis. Compare alternative hypothesis.

observational study. A study in which subjects (or external forces) select themselves into groups; investigators then compare the outcomes for the different groups. For example, studies of smoking are generally observational. Subjects decide whether or not to smoke; the investigators compare the death rate for smokers to the death rate for nonsmokers. In an observational study, the groups may differ in important ways that the investigators do not notice; controlled

experiments minimize this problem. The critical distinction is that in a controlled experiment, the investigators intervene to manipulate the circumstances of the subjects; in an observational study, the investigators are passive observers. (Of course, running a good observational study is hard work, and may be quite useful.) Compare confounding variable; controlled experiment.

observed significance level. A synonym for *p*-value. See significance test. Compare fixed significance level.

odds. The probability that an event will occur divided by the probability that it will not. For example, if the chance of rain tomorrow is $2/3$, then the odds of rain are $(2/3)/(1/3) = 2/1$, or 2 to 1; the odds against rain are 1 to 2.

odds ratio. A measure of association, often used in epidemiology. For example, if 10% of all people exposed to a chemical develop a disease, compared to 5% of people who are not exposed, then the odds of the disease in the exposed group are $10/90 = 1/9$, compared to $5/95 = 1/19$ in the unexposed group. The odds ratio is $(1/9)/(1/19) = 19/9 = 2.1$. An odds ratio of 1 indicates no association. Compare relative risk.

one-sided hypothesis; one-tailed hypothesis. Excludes the possibility that a parameter could be, for example, less than the value asserted in the null hypothesis. A one-sided hypothesis leads to a one-sided (or one-tailed) test. See significance test; statistical hypothesis. Compare two-sided hypothesis.

one-sided test; one-tailed test. See one-sided hypothesis.

overfitting. Occurs when a statistical model, especially one with many free parameters, fits the data used to estimate the model extremely well while generalizing poorly to other data for which it is intended.

outcome variable. See dependent variable.

outlier. An observation that is far removed from the bulk of the data. Outliers may indicate faulty measurements, and they may exert undue influence on summary statistics, such as the mean or the correlation coefficient.

***p*-value.** Result from a statistical test. The probability of getting, just by chance, a test statistic as large as or larger than the observed value. Large *p*-values are consistent with the null hypothesis; small *p*-values undermine the null hypothesis. However, *p* does not give the probability that the null hypothesis is true. The *p*-value is a useful measure of compatibility with the null hypothesis. If *p* is smaller than 5%, the result is often said to be statistically significant. If *p* is smaller than 1%, the result may be described as highly significant. Strict reliance on such cutoffs is not recommended as a best practice. The *p*-value is also called the observed significance level. See significance test; statistical hypothesis.

parameter. A numerical characteristic of a population or a model. See probability model.

percentile. To get the percentiles of a dataset, array the data from the smallest value to the largest. As an example of a percentile, consider the 90th percentile: 90% of the values fall below the 90th percentile and 10% are above. (To be very precise: at least 90% of the data are at the 90th percentile or below; at least 10% of the data are at the 90th percentile or above.) The 50th percentile is the median: 50% of the values fall below the median, and 50% are above.

placebo. See double-blind experiment.

point estimate. An estimate of the value of a quantity expressed as a single number. See estimator. Compare confidence interval; interval estimate.

Poisson distribution. A limiting case of the binomial distribution, when the number of trials is large and the common probability is small. The parameter of the approximating Poisson distribution is the number of trials times the common probability, which is the expected number of events. When this number is large, the Poisson distribution may be approximated by a normal distribution.

population. Also, universe. All the units of interest to the researcher. Compare sample; sampling frame.

population size. Also, size of population. Number of units in the population.

posterior probability. See Bayes' rule.

power. The probability that a statistical test will reject the null hypothesis. To compute power, one has to fix the size of the test and specify parameter values outside the range given by the null hypothesis. A powerful test has a good chance of detecting an effect when there is an effect to be detected. See beta; significance test. Compare alpha; size; p -value.

practical significance. Substantive importance. Statistical significance does not necessarily establish practical significance. With large samples, small differences can be statistically significant. See significance test.

practice effects. Changes in test scores that result from taking the same test twice in succession, or taking two similar tests one after the other.

predicted value. See residual.

predictive validity. A skills test has predictive validity to the extent that test scores are well correlated with later performance, or more generally with outcomes that the test is intended to predict. See validity. Compare reliability.

predictor. See independent variable.

prior probability. See Bayes' rule.

probability. Chance, on a scale from 0 to 1. Impossibility is represented by 0, certainty by 1. Equivalently, chances may be quoted in percent; 100% corresponds to 1, 5% corresponds to .05, and so forth.

probability density. Describes the probability distribution of a continuous random variable. The chance that the random variable falls in an interval equals the area below the density and above the interval. (However, not all random variables have densities.) See probability distribution; random variable.

probability distribution. Gives probabilities for possible values or ranges of values of a random variable. Often, the distribution is described in terms of a density. See probability density.

probability histogram. See histogram.

probability model. Relates probabilities of outcomes to parameters; also, statistical model. The latter connotes unknown parameters.

probability sample. A sample drawn from a sampling frame by some objective chance mechanism; each unit has a known probability of being sampled. Such samples minimize selection bias but can be expensive to draw.

psychometrics. The study of psychological measurement and testing.

qualitative variable; quantitative variable. A qualitative variable describes nonquantitative features of subjects in a study (e.g., marital status: never married, married, widowed, divorced, separated). A quantitative variable describes numerical features of the subjects (e.g., height, weight, income). This is not a hard-and-fast distinction, because qualitative features may be given numerical codes, as with a dummy variable. Quantitative variables may be classified as discrete or continuous. Concepts such as the mean and the standard deviation apply only to quantitative variables. Compare continuous variable; discrete variable; dummy variable. See variable.

quartile. The 25th or 75th percentile. See percentile. Compare median.

quasi-experiment. See natural experiment.

R-squared (R^2). Measures how well a regression equation fits the data. *R*-squared varies between 0 (no fit) and 1 (perfect fit). *R*-squared does not measure the extent to which underlying assumptions are justified. See regression model. Compare multiple correlation coefficient; standard error of regression.

random error. Sources of error that are random in their effect, like draws made at random from a box. These are reflected in the error term of a statistical model. Some authors refer to random error as chance error or sampling error. See regression model.

random variable. A variable whose possible values occur according to some probability mechanism. For example, if a pair of dice are thrown, the total number of spots is a random variable. The chance of two spots is 1/36, the chance of three spots is 2/36, and so forth; the most likely number is 7, with a chance of 6/36. The expected value of a random variable is the weighted average of the possible values; the weights are the probabilities. The expected value need not be a possible value for the random variable.

randomization. See controlled experiment; randomized controlled experiment.

randomized controlled experiment. A controlled experiment in which subjects are placed into the treatment and control groups at random—as if by a lottery. See controlled experiment. Compare observational study.

range. The difference between the biggest and the smallest values in a batch of numbers.

rate. In an epidemiological study, the number of events, divided by the size of the population; often cross-classified by age and gender. For example, the death rate from heart disease among American men in 2020 was about two per thousand. Among women, the rate was about half that.

regression coefficient. The coefficient of a variable in a regression equation. See regression model.

regression diagnostics. Procedures intended to check whether the assumptions of a regression model are appropriate.

regression equation. See regression model.

regression line. The graph of a (simple) regression equation.

regression model. A regression model attempts to combine the values of certain variables (the independent or explanatory variables) in order to get expected values for another variable (the dependent variable). Sometimes, the phrase “regression model” refers to a probability model for the data; if no qualifications are made, the model will generally be linear, and errors will be assumed independent across observations, with common variance. The coefficients in the linear combination are called regression coefficients; these are parameters. At times, “regression model” refers to an equation (“the regression equation”) estimated from data, typically by least squares.

relative frequency. See frequency.

relative risk. A measure of association used in epidemiology. For example, if 10% of all people exposed to a chemical develop a disease, compared to 5% of people who are not exposed, then the disease occurs twice as frequently among the exposed people: The relative risk is 10% divided by 5% = 2. A relative risk of 1 indicates no association. Compare odds ratio.

reliability. The extent to which a measurement process gives the same results on repeated measurement of the same thing. See repeatability, reproducibility. Compare validity.

repeatability. A type of reliability. With a repeatable procedure, measurements by the same examiner using the same equipment at the same place and time should be close to one another. See within-observer variability.

representative sample. Not a well-defined technical term. A sample judged to fairly represent the population, or a sample drawn by a process likely to give

samples that fairly represent the population, for example, a large probability sample.

reproducibility. A type of reliability. With a reproducible procedure, measurements from different examiners often at different places and times also should be similar. See between-observer variability.

resampling. See bootstrap.

residual. The difference between an actual and a predicted value. The predicted value comes typically from a regression equation, and is better called the fitted value, because there is no real prediction going on. See regression model; independent variable.

response variable. See independent variable.

risk factor. See independent variable.

robust. A statistic or procedure that does not change much when data or assumptions are modified slightly.

sample. A set of units collected for study. Compare population.

sample size. Also, size of sample. The number of units in a sample.

sample weights. See stratified random sample.

sampling distribution. The distribution of the values of a statistic, over all possible samples from a population. For example, suppose a random sample is drawn. Some values of the sample mean are more likely; others are less likely. The sampling distribution specifies the chance that the sample mean will fall in one interval rather than another.

sampling error. A sample is part of a population. When a sample is used to estimate a numerical characteristic of the population, the estimate is likely to differ from the population value because the sample is not a perfect microcosm of the whole. If the estimate is unbiased, the difference between the estimate and the exact value is sampling error. More generally, $\text{estimate} = \text{true value} + \text{bias} + \text{sampling error}$. Sampling error is also called chance error or random error. See standard error. Compare bias; nonsampling error.

sampling frame. A list of units designed to represent the entire population as completely as possible. The sample is drawn from the frame.

sampling interval. See systematic sample.

scatter diagram. Also, scatterplot; scattergram. A graph showing the relationship between two variables in a study. Each dot represents one unit from the study, e.g., one subject. One variable is plotted along the horizontal axis, the other variable is plotted along the vertical axis. A scatter diagram is homoscedastic when the spread is more or less the same inside any vertical strip. If the spread changes from one strip to another, the diagram is heteroscedastic.

selection bias. Systematic error due to nonrandom selection of subjects for study.

sensitivity. The probability that a test for the presence of a condition will give a positive result given that the condition is present. Sensitivity is analogous to the power of a statistical test. Compare specificity.

sensitivity analysis. Analyzing data in different ways to see how results depend on methods or assumptions.

sign test. A statistical test based on counting and the binomial distribution. For example, a Finnish study of twins found 22 monozygotic twin pairs where 1 twin smoked, 1 did not, and at least 1 of the twins had died. That sets up a race to death. In 17 cases, the smoker died first; in 5 cases, the nonsmoker died first. The null hypothesis is that smoking does not affect time to death, so the chances are 50–50 for the smoker to die first. On the null hypothesis, the chance that the smoker will win the race 17 or more times out of 22 is 8/1000. That is the p -value. The p -value can be computed from the binomial distribution.

significance level. See fixed significance level; p -value.

significance test. Testing for significance can refer to a hypothesis test performed by computing a p -value and declaring that the test statistic is significant if this p -value is less than or equal to some level (α). Significance testing also can denote just computing the statistic and its p -value to see whether the p -value is large or small; no formal test with a preset significance level is involved. The idea is to see whether the data conform to the predictions of the null hypothesis. Generally, a large test statistic goes with a small p -value; and small p -values would undermine the null hypothesis.

significant. See p -value; practical significance; significance test.

simple random sample. A simple random sample of size n from a population of size N is one in which each set of n units in the sampling frame has the same chance of being chosen as the sample. The investigators take a unit at random (as if by lottery), set it aside, take another at random from what is left, and so forth.

simple regression. A regression equation that includes only one independent variable. Compare multiple regression.

size. A synonym for alpha (α).

skip factor. See systematic sample.

specificity. The probability that a test for the presence of a condition will give a negative result given that the condition is not present. Specificity is analogous to $1-\alpha$, where α is the significance level of a statistical test. Compare sensitivity.

spurious correlation. When two variables are correlated, one is not necessarily the cause of the other. The vocabulary and shoe size of children in elementary school, for example, are correlated—but learning more words will not make the feet grow. Such noncausal correlations are said to be spurious. (Originally, the term seems to have been applied to the correlation between two rates with

the same denominator: Even if the numerators are unrelated, the common denominator will create some association.) Compare confounding variable.

standard deviation (SD). Indicates how far a typical element deviates from the average. For example, in round numbers, the average height of women age 18 and over in the United States is 5 feet 4 inches. However, few women are exactly average; most will deviate from average, at least by a little. The SD is sort of an average deviation from average. For the height distribution, the SD is 3 inches. The height of a typical woman is around 5 feet 4 inches, but is off that average value by something like 3 inches. For distributions that follow the normal curve, about 68% of the elements are in the range from 1 SD below the average to 1 SD above the average. Thus, about 68% of women have heights in the range 5 feet 1 inch to 5 feet 7 inches. Deviations from the average that exceed 3 or 4 SDs are extremely unusual. Many authors use standard deviation to also mean standard error. See standard error.

standard error (SE). Indicates the likely size of the sampling error in an estimate. Many authors use the term standard deviation instead of standard error. Compare expected value; standard deviation.

standard error of regression. Indicates how actual values differ (in some average sense) from the fitted values in a regression model. See regression model; residual. Compare *R*-squared.

standardization. See standardized variable.

standardized variable. A variable transformed to have mean zero and variance one. This involves two steps: (1) subtract the mean, (2) divide by the standard deviation.

statistic. A number that summarizes data. A statistic refers to a sample; a parameter or a true value refers to a population or a probability model.

statistical controls. Procedures that try to filter out the effects of confounding variables on non-experimental data, for example, by adjusting through statistical procedures such as stratification or multiple regression. See multiple regression; confounding variable; observational study. Compare controlled experiment.

statistical dependence. See dependence.

statistical hypothesis. Generally, a statement about parameters in a probability model for the data. The null hypothesis may assert that certain parameters have specified values or fall in specified ranges; the alternative hypothesis would specify other values or ranges. The null hypothesis is tested against the data with a test statistic; the null hypothesis may be rejected if there is a statistically significant difference between the data and the predictions of the null hypothesis. Typically, the investigator seeks to demonstrate the alternative hypothesis; the null hypothesis would explain the findings as a result of mere chance, and the investigator uses a significance or hypothesis test to rule out that possibility.

statistical independence. See independence.

statistical model. See probability model.

statistical significance. See p -value.

statistical test. See significance test.

stratified random sample. A type of probability sample. The researcher divides the population into relatively homogeneous groups called “strata” and draws a random sample separately from each stratum. Dividing the population into strata is called “stratification.” Often the sampling fraction will vary from stratum to stratum. Then sampling weights should be used to extrapolate from the sample to the population. For example, if 1 unit in 10 is sampled from stratum A while 1 unit in 100 is sampled from stratum B, then each unit drawn from A counts as 10, and each unit drawn from B counts as 100. The first kind of unit has weight 10; the second has weight 100.

stratification. See independent variable; stratified random sample.

study validity. See validity.

subjectivist. See Bayesian inference.

systematic error. See bias.

systematic sample. Also, list sample. The elements of the population are numbered consecutively as 1, 2, 3, The investigators choose a starting point and a “sampling interval” or “skip factor” k . Then, every k th element is selected into the sample. If the starting point is 1 and $k = 10$, for example, the sample would consist of items 1, 11, 21, Sometimes the starting point is chosen at random from 1 to k : this is a random-start systematic sample.

t -statistic. A test statistic, used to make the t -test. The t -statistic indicates how far away an estimate is from its expected value, relative to the standard error. The expected value is computed using the null hypothesis that is being tested. With a large sample, a t -statistic larger than 2 or 3 in absolute value makes the null hypothesis rather implausible—the estimate is too many standard errors away from its expected value. See statistical hypothesis; significance test; t -test.

t -test. A statistical test based on the t -statistic. Large t -statistics are beyond the usual range of sampling error. The t -test arises when testing the null hypothesis that the average of a population equals a given value when the population is known to be normally distributed. For small samples, the t -statistic follows Student’s t -distribution (when the null hypothesis holds) rather than the normal curve; larger values of t are required to achieve significance with small samples. The relevant t -distribution depends on the number of degrees of freedom, which in this context equals the sample size minus one. A t -test is not appropriate for small samples drawn from a population that is not normal. See hypothesis test; p -value; significance test; statistical hypothesis.

test statistic. A statistic used to judge whether data conform to the null hypothesis. The parameters of a probability model determine expected values for the data; differences between expected values and observed values are measured by a test statistic. Such test statistics include the chi-squared statistic (χ^2) and the t -statistic. Generally, small values of the test statistic are consistent with the null hypothesis; large values lead to rejection. See hypothesis test; p -value; statistical hypothesis; t -statistic.

time series. A series of data collected over time; for example, the Gross National Product of the United States from 1945 to 2020.

treatment group. See controlled experiment.

two-sided hypothesis; two-tailed hypothesis. An alternative hypothesis asserting that the values of a parameter are different from—either greater than or less than—the value asserted in the null hypothesis. A two-sided alternative hypothesis suggests a two-sided (or two-tailed) test. See significance test; statistical hypothesis. Compare one-sided hypothesis.

two-sided test; two-tailed test. See two-sided hypothesis.

Type I error. A statistical test makes a Type I error when (1) the null hypothesis is true and (2) the test rejects the null hypothesis, i.e., there is a false positive. For example, a study of two groups may show some difference between samples from each group, even when there is no difference in the population. When a statistical test deems the difference to be significant in this situation, it makes a Type I error. See significance test; statistical hypothesis. Compare alpha; Type II error.

Type II error. A statistical test makes a Type II error when (1) the null hypothesis is false and (2) the test fails to reject the null hypothesis, i.e., there is a false negative. For example, there may not be a significant difference between samples from two groups when, in fact, the groups are different. See significance test; statistical hypothesis. Compare beta; Type I error.

unbiased estimator. An estimator that is correct on average over the possible datasets. The estimates have no systematic tendency to be high or low. Compare bias.

uniform distribution. For example, a whole number picked at random from 1 to 100 has a uniform distribution: All values are equally likely. Similarly, a uniform distribution is obtained by picking a real number at random between 0.75 and 3.25: The chance of landing in an interval is proportional to the length of the interval.

validity. Measurement validity is the extent to which an instrument measures what it is supposed to, rather than something else. The validity of a standardized test is often indicated by the correlation coefficient between the test scores and some outcome measure (the criterion variable). See content validity; differential validity; predictive validity. Compare reliability.

Study validity is the extent to which results from a study can be relied upon. Study validity has two aspects, internal and external. A study has high internal validity when its conclusions hold under the particular circumstances of the study. A study has high external validity when its results are generalizable. For example, a well-executed randomized controlled double-blind experiment performed on an unusual study population will have high internal validity because the design is good; but its external validity will be debatable because the study population is unusual.

Validity is used also in its ordinary sense: assumptions are valid when they hold true for the situation at hand.

variable. A property of units in a study, which varies from one unit to another—for example, in a study of households, household income; in a study of people, employment status (employed, unemployed, not in labor force).

variance. The square of the standard deviation. Compare standard error; covariance.

weights. See stratified random sample.

within-observer variability. Differences that occur when an observer measures the same thing twice, or measures two things that are virtually the same. Compare between-observer variability.

z -statistic. A test statistic, used to perform the z -test. The z -statistic indicates how far away an estimate is from its expected value, relative to the standard error. The expected value is computed using the null hypothesis that is being tested. The distinction between the z -statistic and the t -statistic is that the z -statistic requires that the population standard deviation be known (which is not generally the case). Some writers refer to the t -statistic as a z -statistic when the sample size is large. A z -statistic larger than 2 or 3 in absolute value makes the null hypothesis rather implausible—the estimate is too many standard errors away from its expected value. See hypothesis test; statistical hypothesis; significance test; z -test.

z -test. A statistical test based on the z -statistic. The procedure is closely related to the t -test. The distinction between the z -test and the t -test is that the former requires that the population standard deviation is known (generally not the case). Large z -statistics are beyond the usual range of sampling error. For example, if z is bigger than 1.96, or smaller than -1.96 , then the estimate is statistically significant at the 5% level: such values of z are hard to explain on the basis of sampling error. The scale for z -statistics is tied to areas under the relevant probability distribution curve.

The z -test arises when testing the null hypothesis that the average of a population equals a given value when the population is known to be normally distributed with a specified standard deviation. See p -value; significance test; statistical hypothesis.

References on Statistics and Research Methods

Nontechnical Surveys

- Adam Chilton & Kyle Rozema, *Trial by Numbers: A Lawyer's Guide to Statistical Evidence* (2024).
- David Freedman et al., *Statistics* (4th ed. 2007).
- Darrell Huff, *How to Lie with Statistics* (1993).
- Gregory A. Kimble, *How to Use (and Misuse) Statistics* (1978).
- Ethan Bueno de Mesquita & Anthony Fowler, *Thinking Clearly with Data: A Guide to Quantitative Reasoning and Analysis* (2021).
- David S. Moore & William I. Notz, *Statistics: Concepts and Controversies* (10th ed. 2019).
- Michael Oakes, *Statistical Inference: A Commentary for the Social and Behavioral Sciences* (1986).
- Statistics: A Guide to the Unknown* (Roxy Peck et al. eds., 4th ed. 2005).
- David Spiegelhalter, *The Art of Statistics: Learning from Data* (2019).
- Hans Zeisel, *Say It with Figures* (6th ed. 1985).

General References

- Encyclopedia of Statistical Sciences* (Samuel Kotz et al. eds., 2d ed. 2005).
- The Oxford Handbook of Quantitative Methods* (Todd D. Little ed., 2014).
- Best Practices in Quantitative Methods* (Jason W. Osborne ed., 2008).

Reference Guide on Multiple Regression and Advanced Statistical Models

DANIEL L. RUBINFELD AND DAVID CARD

Daniel L. Rubinfeld, Ph.D., is Robert L. Bridges Professor of Law and Professor of Economics at the University of California, Berkeley, Emeritus and Professor of Law at New York University Law School.

David Card, Ph.D., is Class of 1950 Professor of Economics at the University of California, Berkeley.

CONTENTS

Introduction and Overview, 580

Research Design: Model Specification, 590

 What Is the Specific Question That Is Under Investigation
 by the Expert?, 590

 What Model Should Be Used to Evaluate the Question at Issue?, 590

 Choosing the Dependent Variable, 591

 Choosing the Explanatory Variable That Is Relevant to the Question
 at Issue, 592

 Choosing the Additional Explanatory Variables, 592

 Choosing the Form of the Multiple Regression Model, 596

 Dealing with Endogeneity, 597

 Choosing Methods of Analysis Other Than the Basic Multiple
 Regression Method, 600

 Model Selection, 603

 Matching, 604

 Least absolute shrinkage and selection operator (lasso)
 regression, 605

 Random forest regression, 605

Interpreting Multiple Regression Results, 606

 What Is the Practical, as Opposed to the Statistical, Significance
 of the Regression Results?, 606

 When Should Statistical Tests Be Used?, 607

 What Is the Appropriate Level of Statistical Significance?, 608

 Should Statistical Tests Be One-Tailed or Two-Tailed?, 609

 Are the Regression Results Robust?, 610

- What Evidence Exists That the Explanatory Variable Exerts a Causal Effect on the Dependent Variable and Is Not Affected by Endogeneity?, 611
- To What Extent Are the Explanatory Variables Correlated with Each Other?, 611
- How Is the Sample Used in the Regression Model Defined, 613
- Are There Problems with Statistical Inference Owing to Nonindependent Errors?, 613
- To What Extent Are the Regression Results Sensitive to Individual Data Points?, 615
- To What Extent Are the Data Subject to Measurement Error?, 616
- Causal Analysis and Research Designs, 617
 - Randomized Controlled Trials, 619
 - Pre/Post Design, 622
 - Difference-in-Differences Design, 624
 - Generalized Difference-in-Differences Designs, 625
 - Instrumental Variables Design, 626
- More Advanced Regression Methods, 633
 - Fixed-Effects and Random-Effects Regression Models, 633
 - Methods for Estimating Regression Models, 635
- The Expert, 636
 - Who Should Be Qualified as an Expert?, 637
 - Should the Court Appoint a Neutral Expert?, 637
- Presentation of Statistical Evidence, 639
 - Which Database Information and Analytical Procedures Will Aid in Resolving Disputes Over Statistical Studies?, 640
- Appendix: The Basics of Multiple Regression, 641
 - Introduction, 641
 - Multiple Regression Model, 645
 - Specifying the Regression Model, 647
 - Fitted Regression Line, 647
 - Regression Residuals, 649
 - Interaction Terms, 649
 - Interpreting Regression Results, 650
 - The Problem of Omitted Variables, 651
 - Causal Modeling, 653
 - Pre/Post Comparisons, 653
 - Difference in Difference, 656
 - Instrumental Variables, 659
 - Determining the Precision of the Regression Results, 661
 - Standard Errors of the Coefficients and *t*-Statistics, 661
 - Goodness of Fit, 664
 - Sensitivity of Least-Squares Regression Results, 666
 - A Hypothetical Example, 668

Reference Guide on Multiple Regression and Advanced Statistical Models

Glossary of Terms, 671

References on Multiple Regression, 677

References on Causal Analysis, 678

FIGURES

1. Feedback, 598
2. Scatterplot of scores on a job aptitude test relative to job performance rating, 642
3. Correlation between the job aptitude variable and the job performance variable: (a) positive correlation, (b) negative correlation, (c) weak relationship with no correlation, 643
4. Regression line, 645
5. Goodness of fit, 648
6. Regression slope for women's salaries, 651
7. Standard error of the regression (SER), 665
8. Least-squares regression, 667

Introduction and Overview

This reference guide will expand on issues raised in the *Reference Guide on Statistics and Research Methods* discussing how various research designs involving the analysis of more than two variables affect the ability to draw inferences, including those regarding causation. Multiple regression techniques, which are common in litigation, will be the primary focus of this reference guide. Other less common but emerging complex statistical models, such as difference-in-differences designs, will be discussed as they relate to multiple regression and other statistical models. These research designs will be illustrated by actual and hypothetical litigation examples.

Multiple regression analysis is a statistical tool used to understand the relationship between or among two or more variables.¹ Multiple regression involves a variable to be explained—called the dependent variable or outcome variable—and additional explanatory variables that are thought to produce or be associated with changes in the dependent variable.² For example, a multiple regression analysis might estimate the effect of the number of years of work on salary. Salary would be the dependent variable to be explained, while the years of experience would be the explanatory variable.

Multiple regression analysis is sometimes well suited to the analysis of data when there are competing theories proposed to explain the relationship between one variable (or set of variables) and the outcome variable of interest.³ In a case alleging gender discrimination in salaries, for example, a multiple regression analysis could be used to determine whether an average difference in salaries between women and men is attributable wholly or in part to differences between

1. A variable is anything that can take on two or more values (e.g., the daily temperature in Chicago or the salaries of workers at a factory).

2. Explanatory variables in the context of a statistical study are sometimes called independent variables or covariates. See David H. Kaye & Hal S. Stern, *Reference Guide on Statistics and Research Methods*, “Correlation and Regression” in this manual; see also section titled “What Is the Specific Question That Is Under Investigation by the Expert?” below. This reference guide also offers a brief discussion of multiple regression analysis in the section titled “More Advanced Regression Methods” below.

3. Multiple regression is one type of statistical analysis involving several variables. Other types include matching analysis, stratification, analysis of variance, probit analysis, logit analysis, discriminant analysis, and factor analysis.

the two groups in their education and experience.⁴ The employer-defendant might use multiple regression to argue that salary is a function of the employee's education and experience, and the employee-plaintiff might argue that salary is also a function of the individual's sex, even taking account of education and experience. Alternatively, in an antitrust cartel damages case, the plaintiff's expert might utilize multiple regression to evaluate the extent to which the price of a product increased during the period in which the cartel was active, after accounting for costs and other variables unrelated to the cartel. The defendant's expert might use multiple regression to suggest that the plaintiff's expert has omitted several price-determining variables.

More generally, multiple regression may be useful (1) in determining whether a particular effect is present; (2) in measuring the magnitude of a particular effect; and (3) in predicting the value of the dependent variable, but for an intervening event. In a patent infringement case, for example, a multiple regression analysis could be used to estimate (1) whether the behavior of the alleged infringer affected the price of the patented product, (2) the size of the effect, and (3) what the price of the product would have been had the alleged infringement not occurred.

Over the past several decades, the use of multiple regression analysis in court has grown widely. Regression analysis has been used most frequently in cases of

4. Thus, in *Ottaviani v. State University of New York*, 875 F.2d 365, 367 (2d Cir. 1989) (citations omitted), *cert. denied*, 493 U.S. 1021 (1990), the court stated:

In disparate treatment cases involving claims of gender discrimination, plaintiffs typically use multiple regression analysis to isolate the influence of gender on employment decisions relating to a particular job or job benefit, such as salary.

The first step in such a regression analysis is to specify all of the possible "legitimate" (i.e., nondiscriminatory) factors that are likely to significantly affect the dependent variable, and which could account for disparities in the treatment of male and female employees. By identifying those legitimate criteria that affect the decision-making process, individual plaintiffs can make predictions about what job or job benefits similarly situated employees should ideally receive, and then can measure the difference between the predicted treatment and the actual treatment of those employees. If there is a disparity between the predicted and actual outcomes for female employees, plaintiffs in a disparate treatment case can argue that the net "residual" difference represents the unlawful effect of discriminatory animus on the allocation of jobs or job benefits.

sex and race discrimination,⁵ antitrust violations,⁶ and cases involving class certification (under Rule 23).⁷ However, there are a range of other applications,

5. Discrimination cases using multiple regression analysis are legion. See, e.g., *Bazemore v. Friday*, 478 U.S. 385 (1986), *on remand*, 848 F.2d 476 (4th Cir. 1988); *Csicseri v. Bowsher*, 862 F. Supp. 547 (D.D.C. 1994) (age discrimination), *aff'd*, 67 F.3d 972 (D.C. Cir. 1995); *EEOC v. Gen. Tel. Co.*, 885 F.2d 575 (9th Cir. 1989), *cert. denied*, 498 U.S. 950 (1990); *Bridgeport Guardians, Inc. v. City of Bridgeport*, 735 F. Supp. 1126 (D. Conn. 1990), *aff'd*, 933 F.2d 1140 (2d Cir.), *cert. denied*, 502 U.S. 924 (1991); *Bickerstaff v. Vassar College*, 196 F.3d 435, 448–49 (2d Cir. 1999) (sex discrimination); *McReynolds v. Sodexho Marriott*, 349 F. Supp. 2d 1 (D.D.C. 2004) (race discrimination); *Hnot v. Willis Grp. Holdings Ltd.*, 228 F.R.D. 476 (S.D.N.Y. 2005) (gender discrimination); *Carpenter v. Boeing Co.*, 456 F.3d 1183 (10th Cir. 2006) (sex discrimination); *Coward v. ADT Sec. Sys., Inc.*, 140 F.3d 271, 274–75 (D.C. Cir. 1998); *Smith v. Va. Commonwealth Univ.*, 84 F.3d 672 (4th Cir. 1996) (*en banc*); *Hemmings v. Tidyman's Inc.*, 285 F.3d 1174, 1184–86 (9th Cir. 2002); *Mehus v. Emporia State Univ.*, 222 F.R.D. 455 (D. Kan. 2004) (sex discrimination); *Gutierrez v. Johnson & Johnson*, 467 F. Supp. 2d 403 (D.N.J. 2006) (race discrimination); *Morgan v. United Parcel Serv.*, 380 F.3d 459 (8th Cir. 2004) (racial discrimination); *Students for Fair Admissions, Inc. v. Univ. of N.C.*, 567 F. Supp. 3d 580 (M.D.N.C. 2021), *cert. granted*, 142 S. Ct. 896 (2022) (race discrimination in higher education); *City of Oakland v. Wells Fargo & Co.*, 972 F.3d 1112 (9th Cir. 2020), *vacated*, 993 F.3d 1077 (9th Cir. 2021) (racial housing discrimination); *Moussouris v. Microsoft Corp.*, 311 F. Supp. 3d 1223 (W.D. Wash. 2018) (sex discrimination); *Wal-Mart Stores, Inc. v. Dukes*, 564 U.S. 338 (2011) (sex discrimination); *Chen-Oster v. Goldman, Sachs & Co.*, 114 F. Supp. 3d 110 (S.D.N.Y. 2015) (sex discrimination); *Spencer v. Va. State Univ.*, 919 F.3d 199 (E.D. Va. 2019) (sex discrimination). See also Keith N. Hylton & Vincent D. Rougeau, *Lending Discrimination: Economic Theory, Econometric Evidence, and the Community Reinvestment Act*, 85 Geo. L.J. 237, 238 (1996) (“regression analysis is probably the best empirical tool for uncovering discrimination”).

6. E.g., *United States v. Brown Univ.*, 805 F. Supp. 288 (E.D. Pa. 1992) (price fixing of college scholarships), *rev'd*, 5 F.3d 658 (3d Cir. 1993); *Petruzzi's IGA Supermarkets, Inc. v. Darling-Delaware Co.*, 998 F.2d 1224 (3d Cir.), *cert. denied*, 510 U.S. 994 (1993); *Ohio ex rel. Montgomery v. Louis Trauth Dairy, Inc.*, 925 F. Supp. 1247 (S.D. Ohio 1996); *In re Chicken Antitrust Litig.*, 560 F. Supp. 963, 993 (N.D. Ga. 1980); *New York v. Kraft Gen. Foods, Inc.*, 926 F. Supp. 321 (S.D.N.Y. 1995); *Freeland v. AT&T Corp.*, 238 F.R.D. 130 (S.D.N.Y. 2006); *In re Pressure Sensitive Labelstock Antitrust Litig.*, 566 F. Supp. 2d 363 (M.D. Pa., 2008); *In re Linerboard Antitrust Litig.*, 497 F. Supp. 2d 666 (E.D. Pa. 2007) (price fixing by manufacturers of corrugated boards and boxes); *In re Polypropylene Carpet Antitrust Litig.*, 93 F. Supp. 2d 1348 (N.D. Ga. 2000); *In re OSB Antitrust Litig.*, No. 06–826, 2007 WL 2253418 (E.D. Pa. Aug. 3, 2007) (price fixing of Oriented Strand Board, also known as “waferboard”); *In re TFT-LCD (Flat Panel) Antitrust Litig.*, 267 F.R.D. 583 (N.D. Cal. 2010); *In re Urethane Antitrust Litig.*, 768 F.3d 1245 (10th Cir. 2014); *In re Southeastern Milk Antitrust Litig.*, 739 F.3d 262 (6th Cir. 2014); *In re Disposable Contact Lens Antitrust Litig.*, 329 F.R.D. 336 (M.D. Fla. 2018); *In re Broiler Chicken Antitrust Litig.*, No. 16–C-8637, 2022 WL 1720468 (N.D. Ill. May 27, 2022).

For a broad overview of the use of regression methods in antitrust, see *ABA Antitrust, Econometrics: Legal, Practical and Technical Issues* (John Harkrider & Daniel Rubinfeld eds., 2005). See also Jerry Hausman et al., *Competitive Analysis with Differentiated Products*, 34 *Annales D'Économie et de Statistique* 159 (1994), <https://doi.org/10.2307/20075951>; Gregory J. Werden, *Simulating the Effects of Differentiated Products Mergers: A Practical Alternative to Structural Merger Policy*, 5 *Geo. Mason L. Rev.* 363 (1997). For a basic guide, see Michael Cragg et al., *Understanding the Econometric Tools of Antitrust—With No Math!*, 35 *Antitrust Mag.* (2021), <https://perma.cc/P2Z2-Y9CN>.

7. In *Comcast Corp. v. Behrend*, 133 S. Ct. 1426 (2013), the Court found that deficiencies in the plaintiffs' regression model precluded class certification under Rule 23(b)(3). In light of this ruling,

including census undercounts,⁸ voting rights,⁹ the study of the deterrent effect of the death penalty,¹⁰ rate regulation,¹¹ and intellectual property.¹²

however, circuits are split as to the extent to which antitrust plaintiffs must prove that common elements predominate over individual elements. *E.g.*, compare *In re Rail Freight Fuel Surcharge Antitrust Litig.*, 934 F.3d 619 (D.C. Cir. 2019) (finding the number of unharmed plaintiffs defeated predominance) with *Olean Wholesale Grocery Coop., Inc. v. Bumble Bee Foods LLC*, 31 F.4th 651 (9th Cir. 2022) (rehearing en banc) (applying a preponderance of the evidence standard to prerequisites for class certification, but finding that plaintiffs' models satisfied predominance test) and *Kleen Prods. LLC v. Int'l Paper*, 306 F.R.D. 585 (N.D. Ill. 2015) (finding that debate over the expert's multiple regression model "is a merits question that the Court does not need to resolve in order to decide whether to certify the class") *aff'd*, 831 F.3d 919 (7th Cir. 2016), *cert. denied*, 137 S. Ct. 1582 (2017). For a discussion of the use of multiple regression in evaluating class certification, see Bret M. Dickey & Daniel L. Rubinfeld, *Antitrust Class Certification: Towards an Economic Framework*, 66 N.Y.U. Ann. Surv. Am. L. 459 (2010) and John H. Johnson & Gregory K. Leonard, *Economics and the Rigorous Analysis of Class Certification in Antitrust Cases*, 3 J. Competition L. & Econ. 341 (2007), <https://doi.org/10.1093/joclec/nhm009>.

8. See, e.g., *City of New York v. U.S. Dep't of Commerce*, 822 F. Supp. 906 (E.D.N.Y. 1993) (decision of Secretary of Commerce not to adjust the 1990 census was not arbitrary and capricious), *vacated*, 34 F.3d 1114 (2d Cir. 1994) (applying heightened scrutiny), *rev'd sub nom.*, *Wisconsin v. City of New York*, 517 U.S. 1 (1996); *Carey v. Klutznick*, 508 F. Supp. 420, 432–33 (S.D.N.Y. 1980) (use of reasonable and scientifically valid statistical survey or sampling procedures to adjust census figures for the differential undercount is constitutionally permissible), *stay granted*, 449 U.S. 1068 (1980), *rev'd on other grounds*, 653 F.2d 732 (2d Cir. 1981), *cert. denied*, 455 U.S. 999 (1982); *Young v. Klutznick*, 497 F. Supp. 1318, 1331 (E.D. Mich. 1980), *rev'd on other grounds*, 652 F.2d 617 (6th Cir. 1981), *cert. denied*, 455 U.S. 939 (1982).

9. Multiple regression analysis was used in suits charging that at-large area-wide voting was instituted to neutralize Black voting strength, in violation of section 2 of the Voting Rights Act, 42 U.S.C. § 1973 (1988). Multiple regression demonstrated that the race of the candidates and that of the electorate were determinants of voting. See *Williams v. Brown*, 446 U.S. 236 (1980); *Rodriguez v. Pataki*, 308 F. Supp. 2d 346, 414 (S.D.N.Y. 2004); *United States v. Vill. of Port Chester*, No. 06 Civ. 15173 (SCR), 2008 U.S. Dist. LEXIS 4914 (S.D.N.Y. Jan. 17, 2008); *Meza v. Galvin*, 322 F. Supp. 2d 52 (D. Mass. 2004) (violation of VRA with regard to Hispanic voters in Boston); *Bone Shirt v. Hazeltine*, 336 F. Supp. 2d 976 (D.S.D. 2004) (violations of VRA with regard to Native American voters in South Dakota); *Georgia v. Ashcroft*, 195 F. Supp. 2d 25 (D.D.C. 2002) (redistricting of Georgia's state and federal legislative districts); *Benavidez v. City of Irving*, 638 F. Supp. 2d 709 (N.D. Tex. 2009) (challenge of city's at-large voting scheme); *Common Cause v. Rucho*, 279 F. Supp. 3d 587 (M.D.N.C. 2018), *rev'd*, 139 S. Ct. 2484 (2019) (challenge to partisan gerrymander); *Rodriguez v. Harris County*, 964 F. Supp. 2d 686 (S.D. Tex. 2013) (racially motivated gerrymandering and vote dilution claims); *Luna v. County of Kern*, 291 F. Supp. 3d 1088 (E.D. Cal. 2018) (racially motivated vote dilution claim). For commentary on statistical issues in voting rights cases, see, e.g., Daniel L. Rubinfeld, *Statistical and Demographic Issues Underlying Voting Rights Cases*, 15 Evaluation Rev. 659 (1991), <https://doi.org/10.1177/0193841X9101500601>; Stephen P. Klein et al., *Ecological Regression Versus the Secret Ballot*, 31 Jurimetrics J. 393 (1991); James W. Loewen & Bernard Grofman, *Recent Developments in Methods Used in Vote Dilution Litigation*, 21 Urb. Law. 589 (1989); Arthur Lupia & Kenneth McCue, *Why the 1980s Measures of Racially Polarized Voting Are Inadequate for the 1990s*, 12 Law & Pol'y 353 (1990), <https://doi.org/10.1111/j.1467-9930.1990.tb00053.x>; D. James Greiner, *Ecological Inference in Voting Rights Act Disputes: Where Are We Now, and Where Do We Want To Be?*, 47 Jurimetrics J. 115 (2007); D. James Greiner, *Re-Solidifying Racial Bloc Voting: Empirics and Legal*

Multiple regression analysis can be a source of valuable scientific testimony in litigation. When used inappropriately, however, regression analysis can confuse important issues while having little, if any, probative value. In *EEOC v. Sears, Roebuck & Co.*,¹³ in which the U.S. Equal Employment Opportunity Commission (EEOC) charged Sears with discrimination against women in hiring practices, the Seventh Circuit acknowledged that “[m]ultiple regression analyses, designed to determine the effect of several independent variables on a dependent

Doctrine in the Melting Pot, 86 Ind. L.J. 447 (2011); Matt Barreto et al., *A Novel Method for Showing Racially Polarized Voting: Bayesian Improved Surname Geocoding*, 46 N.Y.U. Rev. L. & Soc. Change 1 (2022); Eric Slud et al., Ctr. for Stat. Rsch. & Methodology, U.S. Census Bureau, Statistical Methodology (2016) for Voting Rights Act, Section 203 Determinations (2018).

10. See, e.g., *Gregg v. Georgia*, 428 U.S. 153, 184–86 (1976). For critiques of the validity of the deterrence analysis, see Nat’l Rsch. Council, *Deterrence and Incapacitation: Estimating the Effects of Criminal Sanctions on Crime Rates* (Alfred Blumstein et al. eds., 1978), and updated 2012 report, Nat’l Rsch. Council, *Deterrence and the Death Penalty 2* (Daniel S. Nagin & John V. Pepper eds., 2012) (concluding that the research is “not informative” as to any deterrent effect); Richard O. Lempert, *Desert and Deterrence: An Assessment of the Moral Bases of the Case for Capital Punishment*, 79 Mich. L. Rev. 1177 (1981); Hans Zeisel, *The Deterrent Effect of the Death Penalty: Facts v. Faith*, 1976 Sup. Ct. Rev. 317 (1976); and John Donohue & Justin Wolfers, *Uses and Abuses of Statistical Evidence in the Death Penalty Debate*, 58 Stan. L. Rev. 791 (2005).

11. See, e.g., *Time Warner Entertainment Co., L.P. v. FCC*, 56 F.3d 151 (D.C. Cir. 1995) (challenge to FCC’s application of multiple regression analysis to set cable rates), *cert. denied*, 516 U.S. 1112 (1996); *Appalachian Power Co. v. EPA*, 135 F.3d 791 (D.C. Cir. 1998) (challenging the EPA’s application of regression analysis to set nitrous oxide emission limits); *Consumers Util. Rate Advocacy Div. v. Ark. PSC*, 99 Ark. App. 228 (Ark. Ct. App. 2007) (challenging an increase in non-gas rates); *Qwest Corp. v. Boyle*, 589 F.3d 985 (8th Cir. 2009) (challenge to the Nebraska Public Service Commission’s telecoms rate setting); *In re Rail Freight Fuel Surcharge Antitrust Litig.*, 725 F.3d 244 (D.C. Cir. 2013) (remanding shipping rate case, writing that the *Behrend* decision interpreted Rule 23 to “command” a hard look at regressions at the class certification stage).

12. See *Polaroid Corp. v. Eastman Kodak Co.*, No. 76-1634-MA, 1990 WL 324105, at *29, *62–63 (D. Mass. Oct. 12, 1990) (damages awarded because of patent infringement), *amended by* No. 76-1634-MA, 1991 WL 4087 (D. Mass. Jan. 11, 1991); *Estate of Vane v. The Fair, Inc.*, 849 F.2d 186, 188 (5th Cir. 1988) (lost profits were the result of copyright infringement), *cert. denied*, 488 U.S. 1008 (1989); *Louis Vuitton Malletier v. Dooney & Bourke, Inc.*, 525 F. Supp. 2d 558, 664 (S.D.N.Y. 2007) (trademark infringement and unfair competition suit); *Stone Brewing Co., LLC v. MillerCoors LLC*, 445 F. Supp. 3d 1113 (S.D. Cal. 2020) (utilizing regression analysis in calculating lost profits in trademark infringement case); *Navarro v. P&G*, 515 F. Supp. 3d 718 (S.D. Ohio 2021) (copyright infringement case).

The use of multiple regression analysis to estimate damages has been contemplated in a wide variety of contexts. See, e.g., David Baldus et al., *Improving Judicial Oversight of Jury Damages Assessments: A Proposal for the Comparative Additur/Remittitur Review of Awards for Nonpecuniary Harms and Punitive Damages*, 80 Iowa L. Rev. 1109 (1995); Talcott J. Franklin, *Calculating Damages for Loss of Parental Nurture Through Multiple Regression Analysis*, 52 Wash. & Lee L. Rev. 271 (1995); Roger D. Blair & Amanda Kay Esquibel, *Yardstick Damages in Lost Profit Cases: An Econometric Approach*, 72 Denv. U. L. Rev. 113 (1994); Daniel Rubinfeld, *Quantitative Methods in Antitrust*, in 1 *Issues in Competition Law & Policy* 723 (2008). See also *infra* note 101.

13. 839 F.2d 302 (7th Cir. 1988).

variable, which in this case is hiring, are an accepted and common method of proving disparate treatment claims.”¹⁴ However, the court affirmed the district court’s findings that the “E.E.O.C.’s regression analyses did not ‘accurately reflect Sears’ complex, nondiscriminatory decision-making processes’” and that the “‘E.E.O.C.’s statistical analyses [were] so flawed that they lack[ed] any persuasive value.’”¹⁵ Serious questions also have been raised about the use of multiple regression analysis in census undercount cases and in death penalty cases.¹⁶

The Supreme Court’s rulings in *Daubert* and *Kumho Tire* have encouraged parties to raise questions about the admissibility of multiple regression analyses.¹⁷ Because multiple regression is a well-accepted scientific methodology, courts have frequently admitted testimony based on multiple regression studies, in some cases

14. *Id.* at 324 n.22.

15. *Id.* at 348, 351 (quoting *EEOC v. Sears, Roebuck & Co.*, 628 F. Supp. 1264, 1342, 1352 (N.D. Ill. 1986)). The district court commented specifically on the “severe limits of regression analysis in evaluating complex decision-making processes.” 628 F. Supp. at 1350.

16. See David H. Kaye & Hal S. Stern, *Reference Guide on Statistics and Research Methods*, “Correlation and Regression” and see sections titled “Choosing the Additional Explanatory Variables” and “Choosing the Dependent Variable” below.

17. *Daubert v. Merrill Dow Pharms., Inc.* 509 U.S. 579 (1993); *Kumho Tire Co. v. Carmichael*, 526 U.S. 137, 147 (1999) (expanding the *Daubert* application to nonscientific expert testimony). For example, analysis conducted by PricewaterhouseCoopers on challenges to financial expert witnesses found an approximately eight-fold increase since 2000. PricewaterhouseCoopers, *Daubert Challenges to Financial Experts: A Yearly Study of Trends and Outcomes 4* (2000–2021), <https://perma.cc/38B5-GTTB>. There was a 19% increase of challenges between 2020 and 2021. *Id.* Of those, eighty-nine challenges (33%) resulted in partial or full exclusion of the expert. *Id.*

Circuit courts have developed extensive case law on the issue of applying the *Daubert* standard to expert testimony introduced before the class-action certification stage. Dictum in *Wal-Mart Stores, Inc. v. Dukes*, 564 U.S. 338 (2011), suggested that the district court erred in not applying *Daubert* to expert testimony, and while the Court certified the question in *Comcast Corp. v. Behrend*, 133 S. Ct. 1426 (2013), it did not reach the merits to decide it. In the resulting vacuum, a plurality of circuits held that district courts must submit expert testimony to *Daubert* scrutiny. See *Prantil v. Arkema Inc.*, 986 F.3d 570, 575–76 (5th Cir. 2021) (“if an expert’s opinion would not be admissible at trial, it should not pave the way for certifying a proposed class”); citing *In re Blood Reagents Antitrust Litig.*, 783 F.3d 183, 187 (3d Cir. 2015) (“We join certain of our sister courts to hold that a plaintiff cannot rely on challenged expert testimony, when critical to class certification, to demonstrate conformity with Rule 23 unless the plaintiff also demonstrates, and the trial court finds, that the expert testimony satisfies the standard set out in *Daubert*”); *Sher v. Raytheon Co.*, 419 F. App’x 887, 890–91 (11th Cir. 2011) (“Here the district court refused to conduct a *Daubert*-like critique of the proffered experts’s qualifications. This was error.”); *Am. Honda Motor Co. v. Allen*, 600 F.3d 813, 815–16 (7th Cir. 2010) (“We hold that when an expert’s report or testimony is critical to class certification, as it is here, . . . a district court must conclusively rule on any challenge to the expert’s qualifications or submissions prior to ruling on a class certification motion.”); *Grodzitsky v. Am. Honda Motor Co.*, 957 F.3d 979, 984 (9th Cir. 2020) (“[I]n evaluating challenged expert testimony in support of class certification, a district court should evaluate admissibility under the standard set forth in *Daubert*.”).

over the strong objection of one of the parties.¹⁸ On some occasions courts have excluded expert testimony because of a failure to utilize a multiple regression methodology.¹⁹ On other occasions, courts have rejected regression studies that did not have an adequate foundation or research design with respect to the issues at hand.²⁰

In interpreting the results of a multiple regression analysis, it is important to distinguish between correlation and causality. Two variables are correlated—that is, associated with each other—when the events associated with the variables occur more frequently together than one would expect by chance. For example, if higher salaries are associated with a greater number of years of work experience, and lower salaries are associated with fewer years of experience, there is a positive correlation between salary and number of years of work experience. However, if higher salaries are associated with less experience, and lower salaries are associated with more experience, there is a negative correlation between the two variables.

A correlation between two variables does not necessarily imply that one event causes the second. One common explanation for situations where two variables are correlated but there is no causal connection is spurious correlation.²¹ Spurious correlation arises when two variables move together because they are both caused by a third, unexamined variable. For example, there might be a negative correlation between the age of certain skilled employees of a computer company and their salaries. One should not conclude from this correlation that the employer has necessarily discriminated against the employees based on their age. A third, unexamined variable, such as the level of the employees' technological skills, could explain differences in productivity and, consequently, differences in salary.²² Or

18. See *Newport Ltd. v. Sears, Roebuck & Co.*, Civ. Action No. 86-2319 Section “K,” 1995 U.S. Dist. LEXIS 7652 (E.D. La. May 26, 1995). See also *Petruzzi's IGA Supermarkets, Inc. v. Darling-Delaware Co.*, 998 F.2d 1224, 1240–47 (3d Cir.), cert. denied, 510 U.S. 994 (1993) (finding that the district court abused its discretion in excluding multiple regression-based testimony and reversing the grant of summary judgment to two defendants).

19. See, e.g., *In re Exec. Telecard Ltd. Sec. Litig.*, 979 F. Supp. 1021 (S.D.N.Y. 1997); but see *United States v. Valencia*, 600 F.3d 389 (5th Cir. 2010) (ruling that the lack of a regression analysis was a matter of weight, not admissibility, of expert testimony).

20. See *City of Tuscaloosa v. Harcros Chems., Inc.*, 158 F.3d 548 (11th Cir. 1998), in which the court ruled plaintiffs' regression-based expert testimony inadmissible and granted summary judgment to the defendants. See also *Am. Booksellers Ass'n v. Barnes & Noble, Inc.*, 135 F. Supp. 2d 1031, 1041 (N.D. Cal. 2001), in which a model was said to contain “too many assumptions and simplifications that are not supported by real-world evidence”; see also *Obrey v. Johnson*, 400 F.3d 691 (9th Cir. 2005).

21. See David H. Kaye & Hal S. Stern, *Reference Guide on Statistics and Research Methods*, “What Inferences Can Be Drawn from the Data?” section, in this manual.

22. See, e.g., *Sheehan v. Daily Racing Form Inc.*, 104 F.3d 940, 942 (7th Cir.) (rejecting plaintiff's age discrimination claim because statistical study showing correlation between age and retention ignored the “more than remote possibility that age was correlated with a legitimate job-related qualification”), cert. denied, 521 U.S. 1104 (1997).

consider a patent infringement case in which increased sales of an allegedly infringing product are associated with a lower price of the patented product.²³ This correlation would be spurious if the two products have their own non-competitive market niches and the lower price is the result of a decline in the production costs of the patented product.

Raising the possibility of a spurious correlation will typically not be enough to dispose of a statistical argument asserting causality. It will normally be necessary to show that the third factor that is alleged to cause the spurious correlation is itself relevant. For example, a statistical showing of a relationship between technological skills and worker productivity might be required in the age discrimination example above.²⁴

In most litigation settings involving statistical analysis, causal relationships are specified within the framework of an underlying causal theory that explains the relationship between the two variables. Even when an appropriate theory has been identified, causality cannot be inferred from the theory alone. One must also look for empirical evidence that a causal relationship exists. Conversely, the fact that two variables are correlated does not guarantee the existence of the causal relationship posited in each theory; it could be that the regression model—an approximation of the underlying causal theory—does not reflect the correct interplay among the explanatory variables. Likewise, the absence of correlation does not guarantee that a causal relationship does not exist. Lack of correlation can occur, even when there is a true causal effect from one variable to another if (1) there are insufficient data, (2) the data are measured inaccurately, (3) the data do not allow multiple causal relationships to be sorted out, or (4) the model is specified wrongly because of the omission of a variable or variables that are related to the variable of interest.

In recent years, new methodologies have broadened our ability to draw causal inferences from a variety of data sources. This reference guide includes an explanation of why the randomized controlled trial (RCT) method is widely accepted in the scientific community as offering the strongest evidence of causal relationships. This reference guide also includes a discussion of a standard framework that delineates the precise meaning of “causality” in an RCT, and the extension

23. In some cases, there are statistical tests that allow one to reject claims of causality. For a brief description of these tests, which were developed by Jerry Hausman, see Robert S. Pindyck & Daniel L. Rubinfeld, *Econometric Models and Economic Forecasts* § 7.5 (4th ed. 1997).

24. See, e.g., *Allen v. Seidman*, 881 F.2d 375 (7th Cir. 1989) (judicial skepticism was raised when the defendant did not submit a logistic regression incorporating an omitted variable; defendant’s attack on statistical comparisons must also include an analysis that demonstrates that the comparisons are flawed). The appropriate requirements for the defendant’s showing of spurious correlation could, in general, depend on the discovery process. See, e.g., *Boykin v. Georgia Pac. Co.*, 706 F.2d 1384 (1983) (criticism of a plaintiff’s analysis for not including omitted factors, when plaintiff considered all information on an application form, was inadequate).

of that framework to other nonexperimental research designs.²⁵ In the case of an RCT, the random assignment of individuals to different subgroups that receive different “treatments” helps ensure that the subgroups are composed of similar individuals.²⁶ In that case, differences in the outcome of interest between the subgroups can be attributed to differences in the assigned treatments—since presumably that is the only factor that varies across otherwise randomly determined groups. In the absence of random assignment, individuals with different characteristics typically choose their own groups (or are assigned to different groups by some other party or process). This self-selection into different groups makes it difficult to determine if the differences across groups are caused by differences in exposure to the explanatory variable of interest or to other differences between the groups.

Occasionally, naturally occurring experiments will arise, in which the ordinary operation of a system results in a close approximation of random assignment in a controlled trial. For example, access to an innovative postconviction job training program may be allocated based on a lottery, thereby approximating the random assignment process of a controlled trial. Comparing employment records of those assigned by lottery to the job training program with records of those not assigned by lottery to the program should allow an assessment of the impact of the job training program on subsequent employment history. Such an assessment will be more accurate than simple comparisons between program participants and nonparticipants in settings where participants are self-selected, since often those who self-select into a program are more highly motivated or otherwise different from those who do not. In such cases it is difficult to sort out the causal effects of the program from differences attributable to the motivation (and other characteristics) of the individuals who selected themselves into the two groups.

In the absence of randomized controlled trials or naturally occurring experiments, complex statistical models utilizing and building on regression methods have been developed to allow for drawing causal inferences. Such models allow individuals to sort themselves into different groups and use various forms of statistical analysis as a means of adjusting for such naturally occurring differences, while still allowing for meaningful assessment of differences across the groups. This reference guide will explore how such statistical analyses and adjustments attempt to ensure that the differences identified by the analyses are because of

25. Randomized *clinical* trials are leading examples of RCTs. Many RCTs, however, are not conducted in a clinical environment. See, e.g., the summary of randomized social experiments in David H. Greenberg & Mark Shroder, *Digest of Social Experiments* (3d ed. 2004).

26. In the experimental literature, the different treatment groups are sometimes referred to as “treatment arms.” In the simplest experiment, one treatment arm receives no treatment (or a placebo treatment), while the other arm receives the treatment of interest, such as a new medicine or a new welfare benefit program.

exposure to different research circumstances and not because of differences arising from the self-classification of individuals.

There is a tension between any attempt to reach conclusions with near certainty and the inherently probabilistic nature of multiple regression analysis. In general, multiple regression allows for the expression of uncertainty in terms of probabilities. The reality that statistical analysis generates probabilities concerning relationships rather than certainty should not be seen as an argument against the use of statistical evidence or, worse, as a reason not to admit that there is uncertainty at all. The only alternative might be to use less reliable anecdotal evidence. Instead, the probabilistic nature of the discipline allows for the expression of different levels of confidence in a set of outcomes, upon which a trier of fact may base a decision.

This reference guide addresses several procedural and methodological issues that are relevant in considering the admissibility of, and the weight to be accorded to, the findings of multiple regression and related advanced statistical analyses. It also suggests some standards of reporting and analysis that an expert presenting analyses might be expected to meet.

A brief overview of the sections of this guide: “Research Design: Model Specification” discusses research design—how the basic regression framework can be used to sort out alternative theories about a case. The guide discusses the importance of choosing the appropriate specification of the regression and regression-related models and raises the issue of whether these methods are appropriate for the case at issue. “Interpreting Multiple Regression Results” accepts the regression framework and concentrates on the interpretation of the multiple regression results from both a statistical and a practical point of view. It emphasizes the distinction between regression results that are statistically significant and results that are meaningful to the trier of fact (i.e., practically significant). It also points to the importance of evaluating the robustness of regression analyses, that is, seeing the extent to which the results are sensitive to changes in the underlying assumptions of the regression model. “Causal Analysis and Research Designs” describes a variety of regression-related methodologies that serve as the foundation for causal analysis. Causal analysis methods allow experts to make inferences about but-for worlds—hypothetical worlds that would exist absent alleged wrongful behavior. “More Advanced Regression Methods” offers a brief overview of fixed-effects and random-effects regression models as well as a discussion of model selection methods.

“The Expert” briefly discusses the qualifications of experts and suggests a potentially useful role for court-appointed neutral experts. “Presentation of Statistical Evidence” emphasizes procedural aspects associated with use of the data underlying regression analyses. It encourages greater pretrial efforts by the parties to attempt to resolve disputes over statistical studies.

Throughout the main body of this reference guide, hypothetical examples are used as illustrations. Moreover, the basic mathematics of multiple regression has been kept to a bare minimum. To achieve that goal, the more formal

description of the multiple regression framework has been placed in the Appendix. The Appendix is self-contained and can be read before or after the text. The Appendix also includes further details with respect to the examples used in the body of this reference guide.

Research Design: Model Specification

Multiple regression allows the testifying expert to choose among alternative theories or hypotheses and assists the expert in distinguishing correlations between variables that are plainly spurious from those that may reflect valid relationships.

What Is the Specific Question That Is Under Investigation by the Expert?

Research begins with a clear formulation of a research question. The data to be collected and analyzed must relate directly to this question; otherwise, appropriate inferences cannot be drawn from the statistical analysis. For example, if the question at issue in a patent infringement case is what price the plaintiff's product would have been but for the sale of the defendant's infringing product, sufficient data must be available to allow the expert to account statistically for important factors that determine the price of the product.

What Model Should Be Used to Evaluate the Question at Issue?

Model specification involves several steps, each of which is fundamental to the success of the research effort. Ideally, a multiple regression analysis builds on a theory that describes the variables to be included in the study. A typical regression model will include one or more dependent variables, each of which is believed to be causally related to a series of explanatory variables. Because we cannot be certain that the explanatory variables are themselves unaffected or independent of the influence of the dependent variable (at least at the point of initial study), the explanatory variables are often termed *covariates*. Covariates are known to have an association with the dependent or outcome variable, but causality remains an open question.

For example, the theory of labor markets might lead one to expect that salaries in an industry are related to workers' education, experience, and training. A belief that there is gender discrimination in setting salaries would lead one to create a model in which the dependent variable is a measure of workers' salaries,

and the list of covariates includes an indicator for female gender in addition to measures of training, experience, and education.

We can imagine an alternative world in which an analysis of discrimination in pay setting (or any other issue) might be accomplished through a “natural experiment,” in which for some reason a group of male and female workers who were known to be equally productive were randomly assigned to a variety of employers in an industry under study and asked to fill positions requiring identical experience and skills. In this design, where any difference in salaries could only be a result of discrimination, it would be possible to draw clear and direct inferences from an analysis of salary data. Unfortunately, the opportunity to analyze natural experiments or conduct randomized controlled trials is rarely available to experts in the context of legal proceedings. In the real world, experts must do their best to interpret the results of the available real-world data, recognizing that it is often impossible to control all factors that might affect worker salaries or other outcomes of interest.²⁷

Models are often characterized in terms of parameters (numerical characteristics of the model). In the labor-market discrimination example, one parameter might reflect the increase in salaries associated with each additional year of prior job experience. Another parameter might reflect the difference in salaries associated with jobs in urban versus nonurban areas. Multiple regression uses a sample, or a selection of data, from the population (all the units of interest) to obtain estimates of the values of the parameters of the model. An estimate associated with a particular explanatory variable is an estimated regression coefficient.

Failure to develop the proper theory, failure to choose the appropriate variables, or failure to choose the correct form of the model can substantially bias the statistical results—that is, create a systematic tendency for an estimate of a model parameter to be too high or too low.

Choosing the Dependent Variable

The variable to be explained, the dependent variable, should be the appropriate variable for analyzing the question at issue.²⁸ Suppose, for example, that pay

27. In the literature on natural and quasi-experiments, the explanatory variables are characterized as “treatments” and the dependent variable as the “outcome.” For a review of natural experiments in the criminal justice arena, see David P. Farrington, *A Short History of Randomized Experiments in Criminology*, 27 *Evaluation Rev.* 218–27 (2003), <https://doi.org/10.1177/0193841X03027003002>.

28. In multiple regression analysis, the dependent variable is often a continuous variable that takes on a range of numerical values (like a person’s salary or a test score). When the dependent variable is categorical, taking on only two or three values, modified forms of multiple regression, such as probit analysis or logit analysis, are appropriate. For an example of the use of the latter, see *EEOC v. Sears, Roebuck & Co.*, 839 F.2d 302, 325 (7th Cir. 1988) (*EEOC* used logit analysis to

discrimination among hourly workers is a concern. One choice for the dependent variable is the hourly wage rate of the employees, while another choice is the annual salary. The distinction is important, because annual salary differences may in part result from differences in hours worked. If the number of hours worked is the product of worker preferences and not discrimination, the hourly wage is a good choice. If the number of hours worked is related to the alleged discrimination, annual salary is the more appropriate dependent variable to choose.²⁹

Choosing the Explanatory Variable That Is Relevant to the Question at Issue

The explanatory variable that allows the evaluation of alternative hypotheses must be chosen appropriately. Thus, in a discrimination case, the variable of interest may be the race or sex of the individual. In an antitrust case, it may be a variable that takes on the value 1 to reflect the presence of the alleged anticompetitive behavior and the value 0 otherwise.³⁰

Choosing the Additional Explanatory Variables

An attempt should be made to identify additional known or hypothesized explanatory variables, some of which are measurable and may support alternative substantive hypotheses that can be accounted for by the regression analysis. For example, in a discrimination case, a measure of the skills of the workers may provide an alternative explanation—lower salaries may have been the result of inadequate skills.³¹

measure the impact of variables such as age, education, job-type experience, and product-line experience on the female percentage of commission hires).

29. In job systems in which annual salaries are tied to grade or step levels, the annual salary corresponding to the job position could be the more appropriate dependent variable.

30. Explanatory variables may vary by type, which will affect the interpretation of the regression results. Thus, some variables may be continuous, and others may be categorical.

31. In *James v. Stockham Valves*, 559 F.2d 310 (5th Cir. 1977), the Court of Appeals rejected the employer's claim that skill level rather than race determined assignment and wage levels, noting the circularity of defendant's argument. In *Ottaviani v. State University of New York*, 679 F. Supp. 288, 306–08 (S.D.N.Y. 1988), *aff'd*, 875 F.2d 365 (2d Cir. 1989), *cert. denied*, 493 U.S. 1021 (1990), the court ruled (at the liability phase of the trial) that the university showed that there was no discrimination in either placement into initial rank or promotions between ranks, and so rank was a proper variable in multiple regression analysis to determine whether women faculty members were treated differently than men. *Cf. Bennett v. Nucor Corp.*, 656 F.3d 802, 817–18 (8th Cir. 2011) (faulting plaintiff's expert for omitting experience and qualification variables).

It is neither realistic nor useful to include all possible variables that might influence the dependent variable in a regression model; some cannot be measured, and others may make little difference.³² If a preliminary analysis shows the unexplained portion of the multiple regression to be unacceptably high, the expert may seek to discover whether some previously undetected variable is missing from the analysis.³³

Failure to include a major explanatory variable that is correlated with the explanatory variable of interest in a regression model is an especially serious concern. Because the parameters of a regression model are selected to maximize the

However, in *Trout v. Garrett*, 780 F. Supp. 1396, 1414 (D.D.C. 1991), the court ruled (in the damage phase of the trial) that the extent of civilian employees' prehire work experience was not an appropriate variable in a regression analysis to compute back pay in employment discrimination. According to the court, including the prehire level would have resulted in a finding of no sex discrimination, despite a contrary conclusion in the liability phase of the action. *Id.* See also *Stuart v. Roache*, 951 F.2d 446 (1st Cir. 1991) (allowing only three years of seniority to be considered as the result of prior discrimination), *cert. denied*, 504 U.S. 913 (1992). Whether a particular variable reflects "legitimate" considerations or itself reflects or incorporates illegitimate biases is a recurring theme in discrimination cases. See, e.g., *Moussouris v. Microsoft Corp.*, 311 F. Supp. 3d 1223, 1238 (W.D. Wash. 2018) (finding as appropriate the exclusion of two variables potentially "tainted" by gender bias). See also *Smith v. Va. Commonwealth Univ.*, 84 F.3d 672, 677 (4th Cir. 1996) (en banc) (suggesting that whether "performance factors" should have been included in a regression analysis was a question of material fact); *id.* at 681–82 (Luttig, J., concurring in part) (suggesting that the failure of the regression analysis to include "performance factors" rendered it so incomplete as to be inadmissible); *id.* at 690–91 (Michael, J., dissenting) (suggesting that the regression analysis properly excluded "performance factors"); see also *Diehl v. Xerox Corp.*, 933 F. Supp. 1157, 1168 (W.D.N.Y. 1996). Other times, the inclusion or exclusion of performance factors as a potentially explanatory variable is a question for the trier of fact. See, e.g., *Chi. Teachers Union, Local 1 v. Bd. of Educ. of Chi.*, No. 12–C–10311, 2020 U.S. Dist. LEXIS 32351, at *19–20 (N.D. Ill. Feb. 25, 2020) ("It is for the trier of fact to determine whether [the expert's] failure to account for academic performance in his regressions renders them less probative."). See also *Crawford v. Newport News Indus. Corp.*, No. 4:14–cv–130, 2017 U.S. Dist. LEXIS 118879 (E.D. Va. July 28, 2017) (excluding expert's report that lacked control variable for "specific job classification"); *Anderson v. Westinghouse Savannah River Co.*, 406 F.3d 248 (4th Cir. 2005) (affirming district court's decision to exclude regression analysis that failed to compare similarly situated workers). A challenge that a model omitted certain "non-discriminatory" variables must be able to find differences themselves to sustain the challenge. See *Buchanan v. Tata Consultancy Servs.*, No. 15–cv–01696–YGR, 2017 U.S. Dist. LEXIS 212170 (N.D. Cal. Dec. 27, 2017).

32. The summary effect of the excluded variables shows up as a random-error term in the regression model, as does any modeling error. See Appendix, below, for details. *But see* David W. Peterson, *Reference Guide on Multiple Regression*, 36 *Jurimetrics J.* 213, 214 n.2 (1996) (review essay) (asserting that "the presumption that the combined effect of the explanatory variables omitted from the model are uncorrelated with the included explanatory variables" is "a knife-edge condition . . . not likely to occur").

33. A very low R -squared (R^2) is one indication of an unexplained portion of the multiple regression model that is unacceptably high. However, the inference that one makes from a particular value of R^2 will depend, of necessity, on the context of the issues and particular datasets that are under study. For reasons discussed in the Appendix, a low R^2 does not necessarily imply a poor model (and vice versa).

ability of the model to predict the outcome variable using only the included variables, such an omission may cause an included variable to be incorrectly credited with an effect caused by the excluded variable.³⁴ Such a situation is called omitted variables bias. In this context, bias is a statistical term referring to the difference between the likely (or expected) parameter value that will be estimated in the (flawed) regression model and the true causal effect of the explanatory variable of interest on the outcome variable. A valid regression model will yield unbiased parameter values.

In general, the existence of omitted variables that are likely to be correlated with both the dependent variable and the key explanatory variable of interest in the analysis (such as gender or race in a discrimination analysis) reduce the probative value of the regression analysis. The importance of omitting a relevant variable depends on the strength of the relationship between the omitted variable and the dependent variable and the strength of the correlation between the omitted variable and the explanatory variable of interest. Other things being equal, the greater the correlation between the omitted variable and the variable of interest, the greater the bias caused by the omission. As a result, the omission of an important variable may lead to inferences made from a regression analysis that are incorrect or do not assist the trier of fact.³⁵

34. Technically, the omission of explanatory variables that are correlated with the variable of interest can cause biased estimates of regression parameters.

35. See *Bazemore v. Friday*, 751 F.2d 662, 671–72 (4th Cir. 1984) (upholding the district court’s refusal to accept a multiple regression analysis as proof of discrimination by a preponderance of the evidence, the court of appeals stated that, although the regression used four variable factors (race, education, tenure, and job title), the failure to use other factors, including pay increases that varied by county, precluded their introduction into evidence), *aff’d in part, vacated in part*, 478 U.S. 385 (1986).

Note, however, that in *Sobel v. Yeshiva University*, 839 F.2d 18, 33, 34 (2d Cir. 1988), *cert. denied*, 490 U.S. 1105 (1989), the court made clear that “a [Title VII] defendant challenging the validity of a multiple regression analysis [has] to make a showing that the factors it contends ought to have been included would weaken the showing of salary disparity made by the analysis” by making a specific attack and “a showing of relevance for each particular variable it contends . . . ought to [be] includ[ed]” in the analysis, rather than by simply attacking the results of the plaintiffs’ proof as inadequate for lack of a given variable. See also *Smith v. Va. Commonwealth Univ.*, 84 F.3d 672 (4th Cir. 1996) (en banc) (finding that whether certain variables should have been included in a regression analysis is a question of fact that precludes summary judgment); *Freeland v. AT&T*, 238 F.R.D. 130, 145 (S.D.N.Y. 2006) (“[o]rordinarily, the failure to include a variable in a regression analysis will affect the probative value of the analysis and not its admissibility”).

Also, in *Bazemore v. Friday*, the Court, declaring that the Fourth Circuit’s view of the evidentiary value of the regression analyses was plainly incorrect, stated that “[n]ormally, failure to include variables will affect the analysis’ probativeness, not its admissibility. Importantly, a regression analysis that includes less than all measurable variables may serve to prove a plaintiff’s case.” 478 U.S. 385, 400 (1986) (footnote omitted). Circuits continue to follow this evidentiary ruling. See *Kurtz v. Costco Wholesale Corp.*, 818 Fed. App’x 57 (2d Cir. 2020) and *Karlo v. Pittsburgh Glass Works, LLC*, 849 F.3d 61 (3d Cir. 2017); *but see In re Scrap Metal Antitrust Litig.*, 527 F.3d

Omitted variables that are not correlated with the variable of interest are, in general, less of a concern, because the parameter that measures the effect of the variable of interest on the dependent variable will be estimated without bias. Suppose, for example, that the effect of a policy introduced by the courts to encourage spouses to pay child support has been tested by randomly choosing some cases to be handled according to current court policies and other cases to be handled according to a new, more stringent policy. The effect of the new policy might be measured in a multiple regression using payment success as the dependent variable and a variable indicating whether the old or new policy was in effect as the explanatory variable (1 if the new program was assigned; 0 if it was not). Failure to include an explanatory variable that reflected the age of the husbands involved in the program would not affect the court's evaluation of the new policy, because men of any given age are as likely to be affected by the old policy as they are the new policy. Randomly applying the two policies to each case has ensured that the omitted age variable is not correlated with the policy variable.

Bias caused by the omission of an important variable that is related to the included variables of interest can be a serious problem.³⁶ Nonetheless, it is possible for the expert to account for bias qualitatively if the expert has knowledge (even if not quantifiable) about the relationship between the omitted variable and the explanatory variable. Suppose, for example, that the plaintiff's expert in a sex discrimination pay case is unable to obtain quantifiable data that reflect the skills necessary for a job, but it is known that, on average, women are more skillful than men. Suppose also that a regression analysis of the wage rate of employees (the dependent variable) on years of experience and a variable reflecting the sex of each employee (the key explanatory variable) suggests that men are paid substantially more than women with the same experience. Because differences in skill levels have not been accounted for, the expert may reasonably conclude that the wage difference measured by the regression is a conservative estimate of the true discriminatory effect on wages.³⁷

The precision of the measure of the effect of a variable of interest on the dependent variable is also important.³⁸ In general, the precision of the estimated

517, 530 (6th Cir. 2008) ("That is not to say that a significant error in application will never go to the admissibility, as opposed to the weight, of the evidence.").

36. See also David H. Kaye & Hal S. Stern, *Reference Guide on Statistics and Research Methods*, "What Inferences Can Be Drawn from the Data?" section, in this manual.

37. The inclusion of potentially cherry-picked skill data can likewise serve as a red flag for courts. In *Moussouris v. Microsoft Corp.*, 311 F. Supp. 3d 1223, 1246 (W.D. Wash. 2018), the court found that by omitting younger, less experienced members of the relevant employee cohort, defendant's expert relied on "unrepresentative and thus insufficient" data and excluded her testimony under Federal Rule of Evidence 702.

38. A more precise estimate of a parameter is an estimate with a smaller standard error. The confidence interval associated with a more precise estimate will be smaller. See Appendix, below, for details.

coefficient in a multiple regression model depends on three factors: (1) the sample size (larger samples give more precise results); (2) the extent to which the model successfully explains the dependent variable (higher explanatory power gives more precise results); and (3) the extent to which the variable of interest varies independently of the other covariates in the model (more independence gives more precise results). Sometimes the inclusion of additional covariates can improve the explanatory power of the model and raise the precision of the estimated coefficient of the variable of interest. Including explanatory variables that are irrelevant (i.e., that do not help increase the explanatory power of the model) will reduce the precision of the estimated coefficients. This can cause concern when the sample size is small, but it is not likely to be of great consequence when the sample size is large.

Choosing the Form of the Multiple Regression Model

Choosing the proper set of variables for a multiple regression model does not complete the modeling exercise. The expert must also choose the proper form of the regression model. The most frequently selected form is the linear regression model allowing separate coefficients for each of the explanatory variables in the model (described in the Appendix). In such a model, the magnitude of the change in the dependent variable that is associated with the change in any of the explanatory variables is the same no matter what the level of the explanatory variables. For example, one additional year of experience might add \$5,000 to salary, regardless of the employee's sex or their previous experience.

In some instances, however, there may be reason to believe that changes in explanatory variables will have differential effects on the dependent variable as the values of the explanatory variables change. In these instances, the expert should consider the use of a more flexible model. Suppose, for example, that having two bathrooms in a house adds twice as much value as having only one bathroom. However, the value of a third bathroom is substantially lower than the value of the first or second bathroom, and even more so for the fourth bathroom. This might be accounted for by having the price of a house be a function of (a) the number of bathrooms and (b) the square of the number of bathrooms. We would expect that the first variable would have a positive effect on price, whereas the second variable would have a negative effect. Failure to account for relationships such as this one can lead to either overstatement or understatement of the effect of a change in the value of an explanatory variable on the value of a dependent variable.

Another source of flexibility is to include interactions among the individual explanatory variables. An interaction variable is the product of two other variables that are included in the multiple regression model. The interaction variable allows the expert to consider the possibility that the effect of a change in one variable on the dependent variable may change as the level of another

explanatory variable changes. For example, in a salary discrimination case, a variable of interest might be the sex of the individual, allowing for the possibility that there are wage differences between men and women. However, a more complete description might also allow for the inclusion of a term that interacts (multiplies) a variable measuring experience with the variable representing the sex of the employee (1 if a female employee; 0 if a male employee). This allows the expert to test whether the sex differential varies with the level of experience. A significant negative estimate of the parameter associated with the sex variable would suggest that inexperienced women are discriminated against, whereas a significant negative estimate of the interaction parameter suggests that the extent of discrimination increases with experience.³⁹

There may be cases where a model includes *both* the independent variable of interest and its interaction with some other variable—as in the previous example where female sex is included as one variable and female sex interacted with experience is included as a second variable. Then, one must account for the parameters for both terms to infer the effect of the variable of interest on any group.

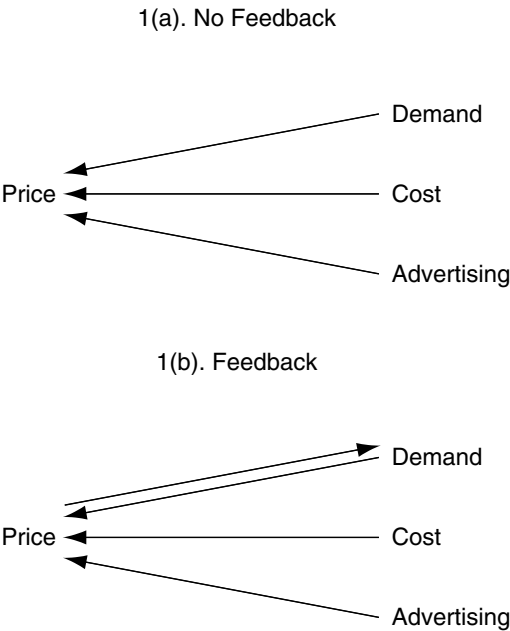
Dealing with Endogeneity

In a multiple regression framework, the expert often assumes that changes in an explanatory variable affect the dependent variable but that changes in the dependent variable do not affect the explanatory variable—that is, there is no feedback.⁴⁰ In cases where there is no feedback, and no spurious correlation owing to unobserved factors that influence both the explanatory and outcome variables, all of the correlation between the explanatory variable and the dependent outcome variable arises from the effect of the former on the latter, and not vice versa. Under

39. For further details concerning interactions, see the Appendix, below. Note that in *Ottaviani v. State Univ. of N.Y.*, 875 F.2d 365, 367 (2d Cir. 1989), *cert. denied*, 493 U.S. 1021 (1990), the defendant relied on a regression model in which a dummy variable reflecting gender appeared as an explanatory variable. The female plaintiff, however, used an alternative approach in which a regression model was developed for men only (the alleged protected group). The salaries of women predicted by this equation were then compared with the actual salaries; a positive difference would, according to the plaintiff, provide evidence of discrimination. For an evaluation of the methodological advantages and disadvantages of this approach, see Joseph L. Gastwirth, *A Clarification of Some Statistical Issues in Watson v. Fort Worth Bank & Trust*, 29 *Jurimetrics J.* 267 (1989).

40. The “no feedback” assumption is not sufficient to ensure that the estimated coefficient on the variable of interest will be unbiased. It must be assumed further that there is no correlation between any omitted variables in the regression equation (as reflected in the error term) and any omitted variables in an equation in which the variable of interest and the outcome variable are reversed (i.e., no spurious correlation). The “no feedback” assumption is especially important in litigation because it is possible for the defendant (if responsible, for example, for price fixing or discrimination) to affect the values of the explanatory variables and thus to bias the usual statistical tests that are used in multiple regression.

Figure 1. Feedback.



these assumptions, the estimated coefficient of the explanatory variable in the multiple regression model represents the causal effect of that variable on the outcome variable. If the “no feedback” assumption is false, or there is a spurious correlation from unobserved factors that affect both variables, the estimated effect of the explanatory variable on the outcome variable is likely to be biased, causing the expert and the trier of fact to reach the wrong conclusion about the true magnitude of the causal effect.

In many situations—particularly those involving the modeling of prices and quantities of a product sold in a market—there is two-way feedback between the outcome variable and the explanatory variable of interest. As a result, if the expert does not take this more complex relationship into account, the regression coefficient on the variable of interest could be either too high or too low. When such two-way feedback is present (or there is a spurious correlation attributable to unobserved factors that affect both variables) the outcome variable and the explanatory variable are said to be “simultaneously determined” (i.e., their relationship

is characterized as one of simultaneity), and the covariate that is believed to be affected by simultaneity is said to be endogenous.⁴¹

Figure 1 illustrates this point. In Figure 1(a), the dependent variable, Price, is explained through a multiple regression framework by three covariate explanatory variables—demand, cost, and advertising—with no feedback. Each of the three covariates is assumed to affect price causally, while price is assumed to have no effect on the three covariates. However, in Figure 1(b), there is feedback, because price affects demand and demand, cost, and advertising affect price. Cost and advertising, however, are not affected by price. In this case both price and demand are jointly determined endogenous variables, that is, each has a causal effect on the other.

As a rule, there are no direct statistical tests for determining the direction of causality; rather, the expert, when asked, should be prepared to defend their assumption based on an understanding of the underlying behavior evidence relating to the businesses or individuals involved.⁴²

Although there is no single approach that is entirely suitable for estimating models when the dependent variable affects one or more of the explanatory variables, one possibility is to drop the questionable variable from the regression to determine whether the variable's exclusion makes a difference. If it does not, the issue becomes moot. Another approach is to expand the multiple regression model by adding one or more equations that explain the relationship between the variable of interest (the explanatory variable) and an outcome variable (the dependent variable). A third approach is to find an instrumental variable—a variable that substantially affects the variable of interest, does not directly affect the outcome variable, and is uncorrelated with any omitted explanatory variables that might affect the dependent variable.

Suppose, for example, that in a salary-based racial discrimination suit the defendant's expert considers employer-evaluated test scores to be an appropriate explanatory variable for the dependent variable, the wage rate. If the plaintiff were to provide information that the employer adjusted the test scores in a manner that penalized the wages of African-American workers, the assumption that wages were determined by test scores alone might be invalid. It might be a totally inappropriate covariate, or it might be endogenous. If the test-score variable is inappropriate, it should be removed from consideration. If endogenous,

41. For discussions of endogeneity problems, see *Conrad v. Jimmy John's Franchise, LLC*, No. 18-CV-00133, 2021 U.S. Dist. LEXIS 84039 (S.D. Ill. Apr. 26, 2021); *In re Nat'l Prescription Opiate Litig.*, No. 1:17-MD-2804, 2019 U.S. Dist. LEXIS 141129 (N.D. Ohio Aug. 20, 2019); and *In re High-Tech Employ. Antitrust Litig.*, 985 F. Supp. 2d 1167 (N.D. Cal. 2013).

42. In settings where each observation in a sample represents a different unit of time—so called “time series” settings—there are statistical formulations of causality that rely on the ordering of events. See Pindyck & Rubinfeld, *supra* note 23, § 9.2. For a more general description of time series analysis, see James H. Stock & Mark W. Watson, *Introduction to Econometrics*, Chapter 15 (2019).

however, the information about the employer's use of the test scores could be translated into a second equation in which a new dependent variable, test score, is related to workers' wages and other variables. A test of the hypothesis that salary and race affect test scores would provide a suitable test of the absence of feedback. Finally, it might be determined that a variable that measures years of experience might be a suitable instrumental variable (i.e., it might be positively correlated with the test scores, yet directly unaffected by the dependent variable salary).⁴³

When a suitable instrumental variable has been found, a more advanced form of multiple regression analysis known as instrumental variables regression will be appropriate.⁴⁴ Intuitively, with this form of regression analysis, the endogenous variable of interest is divided into two parts: (1) a component that is affected by the instrumental variable, but by assumption is not affected by feedback from the dependent variable (or by a spurious correlation); and (2) the remainder. An instrumental variable regression model only uses the part of the endogenous regressor that is affected by the instrumental variable. For more about the use of instrumental variables estimation, see the causal analysis discussion in the section titled "Causal Analysis and Research Designs," below.

Choosing Methods of Analysis Other Than the Basic Multiple Regression Method

There are many multivariate statistical techniques other than the basic multiple regression method that can be useful in legal proceedings. Some statistical methods are appropriate when nonlinearities are important,⁴⁵ while others are appropriate for models in which the dependent variable is discrete, rather than continuous.⁴⁶ Still others have been utilized predominantly to respond to methodological concerns arising in the context of discrimination litigation.⁴⁷

43. Ideally, the instrumental variable should be uncorrelated with any sources of error that might diminish the measured effect of test scores on wages.

44. For examples of this practice in trial, see *United States v. Aetna, Inc.*, 240 F. Supp. 3d 1 (D.D.C. 2017) (antitrust action); see also *In re Domestic Drywall Antitrust Litig.*, 322 F.R.D. 188 (E.D. Pa. 2017).

45. These techniques include, but are not limited to, piecewise linear regression, maximum likelihood estimation of models with nonlinear functional relationships, and autoregressive and moving-average time-series models.

46. For a general discussion of the probit model and its cousin the logit model, see Stock & Watson, *supra* note 42, at Chapter 11.

47. In the analysis of salary discrimination claims, some statisticians have suggested alternative approaches, including urn models (Bruce Levin & Herbert Robbins, *Urn Models for Regression Analysis, with Applications to Employment Discrimination Studies*, 46 Law & Contemp. Probs., 247 (1983)) and, as a means of correcting for measurement errors, reverse regression (Delores A.

It is essential that a valid statistical method be applied to assist with the analysis in each legal proceeding. Therefore, the expert should be prepared to explain why any chosen method, including multiple regression, was more suitable than the alternatives. The following discussion highlights several alternative methods that have proven useful when the dependent variable is discrete rather than continuous.

Suppose that a study has been offered into evidence that purports to evaluate whether there are racial disparities in the imposition of guilty verdicts by juries in criminal cases. One possible goal is to characterize the most important attributes of those cases in which juries find defendants guilty. One covariate might measure the severity of the crime and another might account for the race of the defendant. If the basic regression model is used, predictions generated by the regression model can be interpreted as a measure of the probability that a given defendant will be found to be guilty. Unfortunately, the basic regression model (the “linear probability model”) leaves open the possibility that the predicted probabilities might lie below 0 or above 1—making the interpretation of the regression results nearly impossible.

A common solution to this problem is to utilize a probit or logit model in which the predicted values of the regression model are measures of a probability that lies within the 0 to 1 range. Probit⁴⁸ and logit⁴⁹ models are valuable tools in a wide range of empirical studies, but their results must be interpreted with care. For one thing, the effect of a change in the value of any of the covariates in the model on the probability of an outcome occurring (e.g., a high or a low probability of being found guilty) depends on the baseline probability in the absence of the change. If that probability is very high, a change in the covariate cannot raise the probability much further. Similarly, if the probability is very low, even a large change in the covariate may have a small effect. For example, in a probit or logit model for the probability a student is admitted to a selective college, a given

Conway & Harry V. Roberts, *Reverse Regression, Fairness, and Employment Discrimination*, 1 J. Bus. & Econ. Stat. 75 (1983), <https://doi.org/10.2307/1391775>. But see Arthur S. Goldberger, *Redirecting Reverse Regressions*, 2 J. Bus. & Econ. Stat. 114 (1984); Arlene S. Ash, *The Perverse Logic of Reverse Regression*, in *Statistical Methods in Discrimination Litigation* 85 (D.H. Kaye & Mikel Aickin eds., 1987).

48. Examples of probit analyses span several types of litigation. See *Moussouris v. Microsoft Corp.*, 311 F. Supp. 3d 1223 (W.D. Wash. 2018) (sex discrimination); *Chi. Teachers Union, Local 1 v. Bd. of Educ. of Chi.*, No. 12-C-10311, 2020 U.S. Dist. LEXIS 32351 (N.D. Ill. Feb. 25, 2020) (race discrimination); *In re NCAA Athletic Grant-In-Aid Cap Antitrust Litig.*, No. 4:14-CV-02758, 2017 U.S. Dist. LEXIS 201104 (N.D. Cal. Dec. 6, 2017) (antitrust case); *Georgia v. Ashcroft*, 195 F. Supp. 2d 25 (D.D.C. 2002) (challenge to redistricting plans).

49. Logit models are likewise utilized in a variety of cases. See *W.L. Gore & Assocs. v. C.R. Bard, Inc.*, No. 11-515-LPS-CJB, 2015 U.S. Dist. LEXIS 191654 (D. Del. Sept. 25, 2015) (patent infringement); *Allegre v. Luxottica Retail N. Am.*, 341 F.R.D. 373 (E.D.N.Y. 2022) (false advertising class action); *Laumann v. NHL*, 117 F. Supp. 3d 299 (S.D.N.Y. 2015) (discussing “logit error” in a model in an antitrust violation case).

change in the student's SAT score will have little effect if the student already has an admission probability of 90%, or if the student has a probability of admission close to 0%. But it could have a large effect for students “on the bubble” with roughly a 50% chance of admission. This variation can be addressed by using the estimated model to calculate the change in the probability for each unit in the sample and then averaging the changes to get the average marginal effect of a change in the covariate.

For situations in which the dependent variable takes on only two values, logit and probit models are quite similar. They are also similar in cases where the dependent variable takes on ordered values (like 1, 2, 3, 4, 5)—for example, in situations where the dependent variable is a measure of agreement or sentiment recorded on a 5-point Likert scale (strongly agree, agree, neutral, disagree, strongly disagree).⁵⁰ But in cases where the dependent variable reflects three or more unordered values (e.g., an analysis of consumer demand for sedans, SUVs, and sports cars) a generalization of the logit model known as the multinomial logistic (MNL) model is particularly convenient. Furthermore, if a probabilistic interpretation is useful, the dependent variable can be specified as the logarithm of the odds that a particular choice will be made. In this special case, the coefficients of an estimated logit or MNL model will measure the logarithm of the odds of each of the various choices being made.

Interpretation of the results of probit and logit models can be difficult. To illustrate, assume that the logit model is being used to understand the factors that affect individuals who have recently moved into a city to buy (1) or not to buy (0) a house. Assume also that 60% of the individuals did indeed buy a house. In this case, it would be useful if the expert were asked to answer the following questions:

- a. What are the measured effects of a change in each of the covariates of interest in the model on the probability that there will be a purchase (a 1)?
- b. Are those calculated changes in probability evaluated for every individual in the sample and then averaged? If not, why not?
- c. Do you have a measure of the “goodness of fit” of the logit model? To what extent does the logit model improve in terms of its predictive ability on a model that simply predicts which individuals will buy and which will rent at random (with a probability of purchase being 0.6)?

50. Such models—where the dependent variable involves three or more choices that have a natural ordering—are known as ordered probit and ordered logit regression models.

Model Selection

During testimony, an expert will often offer a regression model that is the result of an iterative model selection process. Depending on the issues in the case, that process might involve varying any or all of the following: (1) the choice of covariates in the model; (2) the form of the model (e.g., measuring the dependent variable as a logarithm or not, including interaction terms or not); (3) the selection of the sample to be used in estimating the model; and (4) the assumptions with respect to the underlying error structure of the model (e.g., accounting for a non-constant error variance or the possibility that errors are correlated over time).⁵¹

A particular concern with model selection in legal settings is that the process of selection was driven in part by the desire of the expert to find a specification in which the coefficient of the variable of interest is either “statistically significant” or not. Such a process is commonly known as *p*-hacking because the statistical significance of an estimate is often expressed by its associated *p*-value.⁵² Informally, this is the probability that the analyst would obtain the estimate in hand, even though the true underlying coefficient (i.e., the one that would be obtained using a very large sample) is zero. As discussed in the section titled “What Is the Appropriate Level of Statistical Significance?” below, it is conventional in scientific work to deem estimated coefficients with *p*-values of less than 5% as “statistically significant.” Thus, an expert who is arguing that there is a “significant effect” of a variable of interest has an incentive to select a specification in which the *p*-value is 5% or smaller. An expert who is arguing the opposite—that the variable of interest has “no significant effect”—has an incentive to select a specification in which the *p*-value is above 5%. *P*-hacking is often suspected when the *p*-value of a selected model is just under or just over 5%, particularly when alternative specifications show much different levels of significance.

Finally, it is sometimes the case that the expert will state explicitly or rely implicitly on regression analyses and tests that were performed either by members of the expert’s team or by other experts or teams. Reliance on the work of others should be reported as part of the expert’s testimony. Otherwise, an expert could evade discovery obligations by having a consulting expert search for a model that the testifying expert then presents as if it were the expert’s own.

The model searching exercise should be distinguished, at least conceptually, from the important process of testing a chosen regression model for robustness.

51. According to Peter Kennedy, “model specification should not blindly follow testing procedures . . . it needs to be a well-thought-out combination of theory and data, and . . . testing procedures used in such specification searches should be designed to minimize the costs of data mining.” Peter E. Kennedy, *Sinning in the Basement: What Are the Rules? The Ten Commandments of Applied Econometrics*, 16 J. Econ. Surveys 569, 578 (2002), <https://doi.org/10.1111/1467-6419.00179>.

52. Megan L. Head et al., *The Extent and Consequences of P-Hacking in Science*, 13 PLoS Biology 1, 15 (2015), <https://doi.org/10.1371/journal.pbio.1002106>.

Robustness checks are applied once a model specification has been selected. The expert applies these checks to report the sensitivity of the reported results to alternative specifications.

The model searching process chosen by an expert may be ad hoc, that is, chosen to be appropriate for the specifics of the issue being studied. However, as computing power has increased, it has become more common for experts to use a variety of algorithmic approaches to estimate model parameters and make predictions. Algorithms can be especially useful when there are many covariates that are potential explanatory variables. Multiple regression is the most basic and most common algorithm, but there are a variety of other model-searching approaches that can benefit from the use of artificial intelligence (AI) methods. They include, but are not limited to, the following:

- Matching
- Least absolute shrinkage and selection operator (lasso) regression
- Random forest regression

Matching

Consider a case where the liability turns on the relationship between an employer's imposition of an anti-female policy (the variable of interest) and whether the individual is promoted or not (the dependent variable). Among all the available observations, some will be more informative than others about the nature of this relationship. To be specific, the most salient information might come by comparing (or matching) women to men of similarly observable characteristics (e.g., age, education, years of employment). There are a variety of algorithmic methods that are designed to find the best match from among all the possible potential matches in the sample of available data.

If one compares the promotion outcomes of women to those of men with similar ages, educational background, and years of experience, a regression model might not need to include those covariates. In its simplest form, the matching approach would measure the impact of the actions of the anti-female policy as the difference in the means of the promotion rate (the dependent variable) of those adversely affected and those not for the reduced sample of matched observations. In most cases there will not be exact matches—hence the need for an algorithmic approach that chooses a subset of the observations likely to be most informative, or more generally weighting the observations from the comparison group in some manner.⁵³

53. Weighting methods are discussed in detail in Nicole Fortin, Thomas Lemieux & Sergio Firpo, *Decomposition Methods in Economics*, in 4 Handbook of Labor Economics, Part A 1–102 (Orley Ashenfelter & David Card eds., 2011).

The matching approach has the advantage that it does not require the expert to specify the form of the underlying regression model—it could be linear or log-linear, for example, or it could include interaction variables. However, matching procedures may end up using a small fraction of the available data from the comparison group, particularly if the group of interest (e.g., female employees) constitutes a small share of the overall sample. In some cases, this loss of sample size may lead to imprecision in the comparison of interest. In addition, matching procedures will generate results of which the significance is not easily tested statistically. Finally, different matching algorithms have different options (sometimes called tuning parameters) that must be set, specifying the tradeoffs to be made in evaluating potential matches and in deciding whether to retain observations when no close match can be found. Since the expert has discretion to choose among these options, it is important for the expert to explain how the choices were made.

Least absolute shrinkage and selection operator (lasso) regression

Lasso can be a useful methodology when there are many covariates that are highly correlated with each other—a situation referred to as multicollinearity. Lasso reduces the impact of multicollinearity by selecting a smaller set of covariates for predictive use in a regression context. If effective, lasso regression can reduce the possible instability of regression models, and it has the potential to improve the prediction process.⁵⁴

Random forest regression

The random forest regression approach is a decision-tree modeling approach that divides the data into two or more parts, searches among possible regression specifications for each part, and then merges the predictions for each part to (ideally) generate a more accurate and stable prediction that would arise if a more basic multiple regression methodology were utilized.⁵⁵ Relying on decision trees can be beneficial when a study involves a large dataset and when one believes that there is a nonlinear and/or highly interactive relationship between the covariates and the dependent variable. The random forest approach is less likely to be of value when the data are

54. Unlike ordinary least squares (OLS), which minimizes the sum of squared residuals, the Lasso estimator minimizes the sum of squares of the values of the coefficients in a model in which all of the variables are specified so that the coefficients are comparable. See Stock & Watson, *supra* note 42, § 14.4. Similarly, a related technique, ridge regression, reduces the number of covariates by utilizing a penalty that increases with the sum of the squared coefficients. See *id.* § 14.3.

55. Of note, the bootstrapping random forest method algorithm creates randomly drawn decision trees from the data, and then averages the results. For a description of the statistical foundation of random forest regression, see Leo Breiman, *Random Forests*, 45 Mach. Learning 5–32 (2001).

in the form of a time series. Time series modeling and analysis often require that one utilize specialized methods to create and then analyze a stationary data series.

Interpreting Multiple Regression Results

Multiple regression results can be interpreted in purely statistical terms, through significance tests, or they can be interpreted in a more practical, nonstatistical manner. Although an evaluation of the practical significance of regression results is almost always relevant in the courtroom, tests of statistical significance are appropriate only in particular circumstances.

What Is the Practical, as Opposed to the Statistical, Significance of the Regression Results?

Practical significance means that the magnitude of the effect being studied is not *de minimis* (i.e., it is sufficiently important for the court to be concerned). For example, if the average weekly wage rate is \$200, a wage differential between men and women of \$2 is likely to be deemed practically insignificant because the differential represents only 1% of the average weekly wage.⁵⁶ That same difference could be statistically significant, however, if a sufficiently large sample of men and women were studied.⁵⁷ The reason is that statistical significance is determined, in part, by the number of observations in the dataset.

Often, results that are practically significant are also statistically significant.⁵⁸ However, it is possible with a large dataset to find statistically significant

56. There is no specific percentage threshold above which a result is practically significant. Practical significance must be evaluated in the context of a particular legal issue. *See also* David H. Kaye & Hal S. Stern, *Reference Guide on Statistics and Research Methods*, “*p*-values, Significance Levels, and Hypothesis Tests” section, in this manual.

57. Practical significance also can apply to the overall credibility of the regression results. Thus in *McCleskey v. Kemp*, 481 U.S. 279 (1987), coefficients on race variables were statistically significant, but the Court declined to find them legally or constitutionally significant.

58. In *Melani v. Board of Higher Education*, 561 F. Supp. 769, 774 (S.D.N.Y. 1983), a Title VII suit was brought against the City University of New York (CUNY) for allegedly discriminating against female instructional staff in the payment of salaries. One approach of the plaintiff’s expert was to use multiple regression analysis. The coefficient on the variable that reflected the sex of the employee was approximately \$1,800 when all years of data were included. Practically (in terms of average wages at the time) and statistically (in terms of a 5% significance test), this result was significant. Thus, the court stated, “Plaintiffs have produced statistically *significant* evidence that women hired as CUNY instructional staff since 1972 received *substantially* lower salaries than similarly

coefficients that are practically insignificant. On the other hand, it is also possible to obtain results that are practically significant but fail to achieve statistical significance (especially when the sample size is small). Suppose for example that an expert undertakes a damages study in a patent-infringement case and predicts but-for sales—what sales would have been had the infringement not occurred—using data that predate the period of alleged infringement. If data limitations are such that only three or four years of pre-infringement sales are known, the difference between but-for sales and actual sales during the period of alleged infringement could be practically significant but statistically insignificant. Alternatively, with only three or four data points, the expert will be unable to detect an effect, even if one exists.

When Should Statistical Tests Be Used?

A test of a specific contention, a hypothesis test, often assists the court in determining whether a violation of the law has occurred in areas in which direct evidence is inaccessible or inconclusive. For example, an expert might use hypothesis tests in race or sex discrimination cases to determine the presence of a discriminatory effect.

Statistical evidence alone can never prove with absolute certainty the worth of any substantive theory. However, by providing evidence contrary to the view that a particular form of discrimination has not occurred, for example, the

qualified men.” *Id.* at 781 (emphasis added). For a related analysis involving multiple comparison, see *Csicseri v. Bousher*, 862 F. Supp. 547, 572 (D.D.C. 1994) (noting that plaintiff’s expert found “statistically significant instances of discrimination” in 2 of 37 statistical comparisons, but suggesting that “2 of 37 amounts to roughly 5% and is hardly indicative of a pattern of discrimination”), *aff’d*, 67 F.3d 972 (D.C. Cir. 1995). See also *Apsley v. Boeing Co.*, 722 F. Supp. 2d 1218 (D. Kan. 2010) (writing that “[s]tatistical significance does not tell us whether the disparity we are observing is meaningful in a practical sense nor what may have caused the disparity” and finding that plaintiffs did not establish a prima facie case that if “forty-eight more people over the age of 40 would have been hired, Plaintiffs’ hiring statistics would not have been statistically significant”). Practical significance continues to be important in other contexts. It often arises in disparate impact cases in the form of a disputed but widely used “four-fifths rule,” an EEOC rule of thumb that if the hire or promotion rate of a minority group is four-fifths that of the comparison group (often white employees), the difference is practically significant. See, e.g., *Hispanic Nat’l L. Enf’t Ass’n NCR v. Prince George’s Cnty.*, 535 F. Supp. 3d 393 (D. Md. 2021) (writing that “pairing [of] a statistical significance test with a practical significance measure is the contemporary standard for demonstrating disparate impact”) (citation omitted); cf. *Jones v. City of Boston*, 752 F.3d 38, 53 (1st Cir. 2014) (concluding that “a plaintiff’s failure to demonstrate practical significance cannot preclude that plaintiff from relying on competent evidence of statistical significance to establish a prima facie case of disparate impact”).

multiple regression approach can aid the trier of fact in assessing the likelihood that discrimination has occurred.⁵⁹

Tests of hypotheses are appropriate in a cross-sectional analysis, in which the data underlying the regression study have been chosen as a sample of a population at a particular point in time, and in a time-series analysis, in which the data being evaluated cover several time periods. In either analysis, the expert may want to evaluate a specific hypothesis, usually relating to a question of liability or to the determination of whether there is measurable impact of an alleged violation. Thus, in a sex discrimination case, an expert may want to evaluate a null hypothesis of no discrimination against the alternative hypothesis that discrimination takes a particular form.⁶⁰ In this case, it is important to realize that rejection of the null hypothesis does not in itself prove legal liability. It is possible to reject the null hypothesis and believe that an alternative explanation other than one involving legal liability accounts for the results.⁶¹

What Is the Appropriate Level of Statistical Significance?

In most scientific work, the level of statistical significance required to reject the null hypothesis (i.e., to obtain a statistically significant result) is set conventionally at 0.05, or 5%.⁶² The significance level measures the probability that the null hypothesis will be rejected incorrectly. In general, the lower the percentage required for statistical significance, the more difficult it is to reject the null hypothesis, and therefore it is less likely that one will err in doing so. Although the 5%

59. See *Int'l Brotherhood of Teamsters v. United States*, 431 U.S. 324 (1977) (the Court inferred discrimination from overwhelming statistical evidence by a preponderance of the evidence); *Ryther v. KARE 11*, 108 F.3d 832, 844 (8th Cir. 1997) (“The plaintiff produced overwhelming evidence as to the elements of a prima facie case, and strong evidence of pretext, which, when considered with indications of age-based animus in [plaintiff’s] work environment, clearly provide sufficient evidence as a matter of law to allow the trier of fact to find intentional discrimination.”); *Paige v. California*, 291 F.3d 1141 (9th Cir. 2002) (allowing plaintiffs to rely on aggregated data to show employment discrimination).

60. Tests are also appropriate when comparing the outcomes of a set of employer decisions with those that would have been obtained had the employer made a different choice from among the available options.

61. See David H. Kaye & Hal S. Stern, *Reference Guide on Statistics and Research Methods*, “Evaluating Hypothesis Tests” in this manual, and see the section titled “Difference-in-Differences Design” below.

62. See, e.g., *Palmer v. Shultz*, 815 F.2d 84, 92 (D.C. Cir. 1987) (“the .05 level of significance . . . [is] certainly sufficient to support an inference of discrimination” (quoting *Segar v. Smith*, 738 F.2d 1249, 1283 (D.C. Cir. 1984), cert. denied, 471 U.S. 1115 (1985))); *United States v. Delaware*, Civil Action No. 01-020-KAJ, 2004 U.S. Dist. LEXIS 4560 (D. Del. Mar. 22, 2004) (stating that .05 is the normal standard chosen).

criterion is typical, reporting of more stringent 1% significance tests or less stringent 10% tests can also provide useful information.

In conducting a statistical test, it is useful to compute an observed significance level, or p -value. The p -value associated with the null hypothesis that a regression coefficient is 0 is the probability that a coefficient of this magnitude or larger could have occurred by chance if the null hypothesis were true. If the p -value were less than or equal to 5%, the expert would reject the null hypothesis in favor of the alternative hypothesis, whereas if the p -value were greater than 5%, the expert would fail to reject the null hypothesis. The use of 1%, 5%, and sometimes 10% levels for determining statistical significance remains a subject of debate. One might argue, for example, that when there is a relatively specific alternative to the null hypothesis, a somewhat lower level of confidence might be appropriate. Otherwise, when the alternative to the null hypothesis involves a vague alternative of “effect,” a high level of confidence (associated with a low significance level, such as 1%) may be appropriate.⁶³

Should Statistical Tests be One-Tailed or Two-Tailed?

When the expert evaluates the null hypothesis that a variable of interest has no linear association with a dependent variable against the alternative hypothesis that there is an association, a two-tailed test, which allows for the effect to be either positive or negative, is usually appropriate. A one-tailed test would usually be applied when the expert believes, perhaps based on other direct evidence presented at trial, that the alternative hypothesis is either positive or negative, but not both. For example, an expert might use a one-tailed test in a patent infringement case if they strongly believe that the effect of the alleged infringement on the price of the infringed product was either zero or negative. (The sales of the infringing product competed with the sales of the infringed product, thereby

63. See, e.g., *Vuyanich v. Republic Nat'l Bank*, 505 F. Supp. 224, 272 (N.D. Tex. 1980) (noting the “arbitrary nature of the adoption of the 5% level of [statistical] significance” to be required in a legal context); *Cook v. Rockwell Int'l Corp.*, 580 F. Supp. 2d 1071 (D. Colo. 2006). Indeed, “courts generally have found that challenges to statistical significance go to the weight, but not the admissibility, of a regression model.” *In re EpiPen (Epinephrine Injection, USP) Mktg., Sales Pracs. and Antitrust Litig.*, No. 17-MD-2785-DDC-TJJ, 2020 U.S. Dist. LEXIS 40788, at *131 (D. Kan. Feb. 27, 2020). See *Kurtz v. Kimberly-Clark Corp.*, 414 F. Supp. 3d 317, 331 (E.D.N.Y. 2019) (“Regressions should not be excluded on the ground that they fail to meet arbitrary thresholds of statistical significance.”); *In re High-Tech Emp. Antitrust Litig.*, No. 11-CV-02509-LHK, 2014 WL 1351040, at *15 (N.D. Cal. Apr. 4, 2014) (“The Court finds that the fact that these two variables are not statistically significant at the 1%, 5%, and 10% levels goes to the weight, not the admissibility of [the] model.”); *EEOC v. Mavis Discount Tire, Inc.*, 129 F. Supp. 3d 90, 110 (S.D.N.Y. 2015) (writing that while courts have rejected a “formal” litmus test for Title VII claims, two or three standard deviations “can be highly probative”) (citation omitted).

lowering the price.) By using a one-tailed test, the expert is in effect stating that without analyzing the data, it would be very surprising if the data pointed in the direct opposite direction to the one posited by the expert.

Because using a one-tailed test produces p -values that are one-half the size of p -values using a two-tailed test, the choice of a one-tailed test makes it easier for the expert to reject a null hypothesis. Correspondingly, the choice of a two-tailed test makes null hypothesis rejection less likely. Because there is some arbitrariness involved in the choice of an alternative hypothesis, courts should avoid relying solely on sharply defined statistical tests.⁶⁴ However, reporting the p -value or a confidence interval should be encouraged because it conveys useful information to the court, whether a null hypothesis is rejected or not.

Are the Regression Results Robust?

The issue of robustness—whether regression results are sensitive to slight modifications in assumptions (e.g., that the data are measured accurately)—is of vital importance. If the assumptions of the regression model are valid, standard statistical tests can be applied. When the assumptions of the model are violated, standard tests can overstate or understate the significance of the results.

The violation of an assumption does not necessarily invalidate a regression analysis, however. The following questions highlight some of the more important assumptions of regression analysis.

64. Courts have shown a preference for two-tailed tests. *See, e.g.,* *Palmer v. Shultz*, 815 F.2d 84, 95–96 (D.C. Cir. 1987) (rejecting the use of one-tailed tests, the court found that because some appellants were claiming over-selection for certain jobs, a two-tailed test was more appropriate in Title VII cases); *Smith v. City of Boston*, 144 F. Supp. 3d 177, 196 (D. Mass. 2015) (observing in a Title VII case, “[t]he weight of the case law appears to favor two-tailed tests”); *Moore v. Summers*, 113 F. Supp. 2d 5, 20 (D.D.C. 2000) (reiterating the preference for a two-tailed test). *See also* *Csicseri v. Bowsher*, 862 F. Supp. 547, 565 (D.D.C. 1994) (finding that although a one-tailed test is “not without merit,” a two-tailed test is preferable); *but see In re EpiPen* (Epinephrine Injection, USP) Mktg., Sales Pracs. and Antitrust Litig., No. 17-MD-2785-DDC-TJJ, 2020 U.S. Dist. LEXIS 40788, at *131 (D. Kan. Feb. 27, 2020) (allowing a model with a one-tailed test to move to the factfinder, writing that the expert’s reasoning was “at worst, plausible”). *See also* David H. Kaye & Hal S. Stern, *Reference Guide on Statistics and Research Methods*, “Evaluating Hypothesis Tests” in this manual, and see also the section titled “Difference-in-Differences Design” below.

What Evidence Exists That the Explanatory Variable Exerts a Causal Effect on the Dependent Variable and Is Not Affected by Endogeneity?

In a multiple regression framework, the expert often assumes that changes in one or more of the explanatory variables affect the dependent variable, but changes in the dependent variable do not affect the explanatory variables (i.e., no feedback, as described previously), nor is there any spurious correlation arising from unmeasured factors that affect both variables. If the causality were reversed, or if there were unmeasured factors leading to spurious correlation, the expert and the trier of fact would likely reach the wrong conclusion from the reported results.

As noted in the section titled “Choosing the Additional Explanatory Variables” above, there are no basic, direct statistical tests for determining the direction of causality. Rather, it may be appropriate to ask (1) whether there is any concern over feedback (or reverse causality) from the outcome variable to the explanatory variable; and (2) whether there is any concern over possible unmeasured factors that affect both variables.

If there is a possibility of feedback from the outcome variable to one or more of the explanatory variables in the regression model, it may be useful to ask the following questions:

- a. How do the results change if the questionable explanatory variable(s) are dropped from the model?
- b. Has the expert developed a model relating the questionable variable to the outcome variable and the other covariates? If so, is there evidence of feedback from the outcome variable to the questionable variable?

To What Extent Are the Explanatory Variables Correlated with Each Other?

It is essential in multiple regression analysis that the explanatory variable of interest not be correlated perfectly with one or more of the other explanatory variables. In essence, with perfect correlation there are two explanations for the same pattern in the data. Suppose, for example, that in a gender discrimination suit, a particular form of job experience is determined to be a valid source of high wages. If all men had the requisite job experience and all women did not, it would be impossible to tell whether wage differentials between men and women resulted from gender discrimination or simply differences in years of experience.

When two or more explanatory variables are correlated perfectly—that is, when there is perfect collinearity—one cannot estimate the regression parameters. The existing dataset does not allow one to distinguish between

alternative competing explanations of the movement in the dependent variable. However, when two or more variables are highly, but not perfectly, correlated—that is, when there is multicollinearity—the regression can be estimated, but some concerns remain. As discussed above, the greater the multicollinearity between two variables, the less precise are the estimates of individual regression parameters, and the less an expert is able to distinguish among competing explanations for the movement in the outcome variable (even though there is no problem in estimating the joint influence of the two variables and all other regression parameters).⁶⁵

Fortunately, the reported regression statistics take into account any multicollinearity that might be present.⁶⁶ However, it is important to note as a corollary that a failure to find a strong relationship between a variable of interest and a dependent variable need not imply that there is no relationship.⁶⁷ A relatively small sample, or even a large sample with substantial multicollinearity, may not provide sufficient information for the expert to determine whether there is a relationship.

65. See *Griggs v. Duke Power Co.*, 401 U.S. 424 (1971) (The court argued that an education requirement was one rationalization of the data, but racial discrimination was another. Putting both race and education in the regression, it would have been asking too much of the data to tell which variable was doing the real work, because education and race were so highly correlated in the market at that time.).

66. See *Denny v. Westfield State College*, 669 F. Supp. 1146, 1149 (D. Mass. 1987) (The court accepted the testimony of one expert that “the presence of multicollinearity would merely tend to *overestimate* the amount of error associated with the estimate. In other words, *p*-values will be artificially higher than they would be if there were no multicollinearity present.”) (emphasis added); *In re High Fructose Corn Syrup Antitrust Litig.*, 295 F.3d 651, 659 (7th Cir. Ill. 2002) (refusing to second-guess district court’s admission of regression analyses that addressed multicollinearity in different ways). Multicollinearity has not been a reason to preclude a model from the factfinder. See *In re Air Cargo Shipping Servs. Antitrust Litig.*, No. 06-MD-1175, 2014 WL 7882100, at *19 (E.D.N.Y. Oct. 15, 2014), *report and recommendation adopted*, 2015 WL 5093503 (E.D.N.Y. July 10, 2015) (“The fact that Kaplan’s model may be tainted by multicollinearity goes purely to its weight, and is not a reason to strike it.”); *In re High-Tech Employee Antitrust Litig.*, No. 11-CV-02509, 2014 WL 1351040, at *21 (N.D. Cal. Apr. 4, 2014) (“Other courts have admitted regressions even in the face of expert disagreement regarding whether collinearity posed a problem. . . . This is not surprising given that the concept of collinearity is not a methodology, but a common phenomenon that results when using the methodology of regression analysis.”); *In re Korean Ramen Antitrust Litig.*, No. 13-CV-04115-WHO, 2017 U.S. Dist. LEXIS 7756 (N.D. Cal. Jan. 19, 2017) (finding a permissible level of multicollinearity and dispute over variable inclusion did not make model unreliable). *But see* *Reed Const. Data Inc. v. McGraw-Hill Cos., Inc.*, 49 F. Supp. 3d 385, 405 (S.D.N.Y. 2014), *aff’d*, 638 F. App’x 43 (2d Cir. 2016) (excluding expert’s method, in part, because of severe multicollinearity).

67. If an explanatory variable of concern and another explanatory variable are highly correlated, dropping the second variable from the regression can be instructive. If the coefficient on the explanatory variable of concern becomes significant, a relationship between the dependent variable and the explanatory variable of concern is suggested.

How Is the Sample Used in the Regression Model Defined?

One potential source of bias in regression modeling arises when the sample used to estimate the model represents only a subsample of the underlying population, and inclusion in the sample is based on a criterion that may be related to the outcome or to the explanatory variables in the proposed regression model. For example, consider an analysis of possible racial disparities in the imposition of guilty verdicts by juries in criminal cases. Such an analysis can only be conducted for cases that go to trial. But if attorneys for nonwhite defendants are aware of potential jury bias, they may recommend a plea bargain unless they believe the case is highly unlikely to yield a guilty verdict despite the race of the defendant. In that situation, the set of cases with nonwhite defendants that appear at trial will tend to be highly selective, and less likely to have a guilty verdict than other cases, leading to a sample selection bias in the estimated effect of defendant race on the probability of a guilty verdict.⁶⁸ Sample selection bias can also arise even when an initial sample is randomly selected, if inclusion in the final sample used in the model estimation depends on the presence of certain information (for example, the availability of complete data on key variables).

The expert should be prepared to explain the derivation of the sample used in estimation. In situations where the estimation sample differs from the underlying population and is not based on a random sample selection rule, the expert should be prepared to defend the sample selection criteria and explain why these criteria do not lead to sample selection bias.

Are There Problems with Statistical Inference Owing to Nonindependent Errors?

Multiple regression analysis yields a set of estimated coefficients that measure the effects of each of the explanatory variables on the outcome variable, holding constant the other explanatory variables. Standard regression modeling software also reports a standard error for each coefficient that can be used to form a *t*-statistic or a confidence interval.⁶⁹ To this point we have emphasized problems such as endogeneity that can arise in interpreting these estimated coefficients from a regression model. But a second set of problems can arise in estimating the standard

68. Sample selection bias is also an issue for comparisons of the outcomes of police searches of motorists, since only motorists who are stopped in the first place are searched. See, e.g., John Knowles, Nicola Persico & Petra Todd, *Racial Bias in Motor Vehicle Searches: Theory and Evidence*, 109 J. Pol. Econ. 203 (2001).

69. For a discussion of confidence intervals, see Stock & Watson, *supra* note 42, at Chapter 3. Loosely speaking, a confidence interval represents an interval of values in which the true value of a regression coefficient falls within some pre-specified probability (where the true value is the estimate that would be obtained from the same model with a very large sample).

errors, even when the coefficient estimates are unbiased. An important assumption underlying the procedure used to estimate these standard errors is that the error terms in the regression model are independent of each other and independent of the explanatory variables.⁷⁰ Independence of the errors is a much stronger assumption than is needed to ensure unbiasedness of the coefficient estimates.

The assumption of independence may be inappropriate in several circumstances. If this strong assumption does not hold and the estimates of the standard errors are themselves biased, the analyst might draw inappropriate conclusions about the confidence that should be attributed to the size of the regression coefficients. In some instances, failure of the assumption makes multiple regression analysis an unsuitable statistical technique; in other instances, modifications or adjustments within the regression framework can be made to accommodate the failure.

The independence assumption may fail, for example, in a study of individual behavior over time, in which an unusually high error value in one period is likely to lead to an unusually high value in the next period. For example, if an economic forecast model underpredicted this year's gross domestic product, the model is likely to underpredict next year's as well; the factor that caused the prediction error (e.g., an incorrect assumption about Federal Reserve policy) is likely to be a continuing source of error in the future.⁷¹

Alternatively, the assumption of independence may fail in a study of a group of firms at a particular point in time in which the error terms for large firms are systematically larger in absolute value than the error terms for small firms.⁷² For example, an analysis of the profitability of firms will include in the error term all the unaccounted-for factors that lead to higher or lower profit. For a large national retail firm with many stores, these omitted factors can lead to errors of plus or minus several million dollars in magnitude. For a small local retailer with one store, however, the range of omitted factors is smaller, and the errors may only be plus or minus a few thousand dollars.

A third possibility is that the dependent variable varies at the individual level, but the explanatory variable of interest varies only at the level of a group. For example, an expert might be viewing the price of a product in an antitrust case as a function of a variable or variables that depend on the marketing channel through which the product is sold (e.g., wholesale or retail). In this case, errors

70. When two variables are independent there is no information in one variable about any feature of the other variable. Independence implies that the variables are uncorrelated, which only requires that there is no information in one variable about the mean value of the other.

71. In this case, the errors in the regression model are said to be serially correlated.

72. This general class of problems, in which the variability in the error term is related to the covariates, is known as conditional heteroscedasticity. This problem also arises when the dependent variable is a binary variable (i.e., with a value of 0 or 1), because the range that the error term can take is limited.

within each of the marketing groups are not likely to be independent. Failure to account for this could cause the expert to overstate the statistical significance of the regression parameters.

In some instances, there are statistical tests that are appropriate for evaluating the assumption that the error terms are independent of each other and of the covariates in the model.⁷³ If the assumption has failed, there are more advanced versions of the procedures to estimate standard errors that can be used to correct for correlation in the errors over time, or for differences in the dispersion in the error component that depend on the covariates, or for dependence between observations from related subgroups of the dataset.⁷⁴ In some cases it is also possible to estimate an alternative version of the regression model that takes into account the dependence of the error terms for different observations.⁷⁵

To What Extent Are the Regression Results Sensitive to Individual Data Points?

Evaluating the robustness of multiple regression results is a complex endeavor. Consequently, there is no agreed upon set of tests for robustness that analysts should apply. In general, it is important to explore the reasons for unusual data points. If the source is an error in recording data, the appropriate corrections can be made. If all the unusual data points have certain characteristics in common (e.g., they all are associated with a supervisor who consistently gives high ratings in an equal pay case), the regression model should be modified appropriately.

73. In a time-series analysis, the correlation of error values over time, the serial correlation, can be tested (in most instances) using several tests, including the Durbin-Watson test. The possibility that some error terms are consistently high in magnitude and others are systematically low—a form of heteroscedasticity—can also be tested in several ways. See, e.g., Pindyck & Rubinfeld, *supra* note 23, at 146–59.

74. When serial correlation and/or heteroscedasticity are present, the standard errors associated with the estimated coefficients must be modified. For a discussion of the use of such “robust” standard errors, see Jeffrey M. Wooldridge, *Introductory Econometrics: A Modern Approach*, Chapter 8 (4th ed. 2009). For a discussion of the treatment of standard errors when the analysis involves panel data (cross-sections and time series), see Stock & Watson, *supra* note 42, at Chapter 10.

75. When serial correlation is present, several closely related statistical methods are appropriate, including generalized differencing (a type of generalized least squares) and maximum likelihood estimation. When heteroscedasticity is the problem, weighted least squares and maximum likelihood estimation are appropriate. See Stock & Watson, *supra* note 42, at Chapter 16. All these techniques are readily available in a variety of statistical computer packages. They also allow one to perform the appropriate statistical tests of the significance of the regression coefficients.

One might think that the sensitivity of regression results can be readily evaluated through an analysis of the regression residuals.⁷⁶ Unfortunately, this is not *always* the case. Given that the basic regression model penalizes estimated prediction errors by the square of the error, an outlier, a highly unusual data point that is more than some appropriate distance from a regression line (even if measured incorrectly), may affect the regression line very substantially, with the result being a relatively low regression residual.

To test to see if this is a problem, one useful diagnostic technique is to determine to what extent the estimated parameter changes as each data point in the regression analysis is dropped from the sample. An influential data point—a point that causes the estimated parameter to change substantially—should be studied further to determine whether mistakes were made in the use of the data or whether important explanatory variables were omitted.⁷⁷

One final cautionary note: What is or is not an outlier is subjective. This leaves the possibility that an expert might point to an outlier or outliers in support of a particular point of view.

To What Extent Are the Data Subject to Measurement Error?

In multiple regression analysis it is assumed that variables are measured accurately.⁷⁸ If there are measurement errors in the dependent variable, estimates of regression parameters will be less precise, but they will not necessarily be biased.⁷⁹ However, if one or more independent variables are measured with error, the corresponding parameter estimates are likely to be biased, typically toward zero (and other coefficient estimates are likely to be biased as well). As the measurement error increases, the estimated parameter associated with the noisily measured

76. A regression residual is the difference between the actual value of the dependent variable and the value that is predicted by the estimated regression model.

77. A more complete and formal treatment of the robustness issue appears in David A. Belsley et al., *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity* 229–44 (1980). For a useful discussion of the detection of outliers and the evaluation of influential data points, see R.D. Cook & S. Weisberg, *Residuals and Influence in Regression* (Monographs on Stat. and Applied Probability No. 18, 1982). For a broad discussion of robust regression methods, see Peer J. Rousseeuw & Annick M. Leroy, *Robust Regression and Outlier Detection* (2003). See generally Peter J. Huber & Elvezio M. Rochetti, *Robust Statistics* (2d ed. 2009).

78. Inaccuracy can occur not only in the precision with which a particular variable is measured, but also in the precision with which the variable to be measured corresponds to the appropriate theoretical construct specified by the regression model.

79. An exception to this rule arises when the dependent variable is dichotomous (or takes on a discrete set of values). See Jerry Hausman, *Mismeasured Variables in Econometric Analysis: Problems from the Right and Problems from the Left*, 15 J. Economic Perspectives 57, 67 (2001), <https://doi.org/10.1257/jep.15.4.57>.

variable will tend toward 0, that is, eventually there will be no relationship with the dependent variable.

It is important for any source of measurement error to be carefully evaluated. In some circumstances, little can be done to correct the measurement-error problem, and the regression results must be interpreted in that light. In other circumstances, however, the expert can correct measurement error by finding a new, more reliable data source. Finally, alternative estimation techniques (using related variables that are measured without error) can be applied to remedy the measurement-error problem in some situations.⁸⁰

Causal Analysis and Research Designs

Multiple regression is a powerful tool for measuring the association between a specific variable or covariate of interest, such as employee gender, and an outcome, such as salary, while holding constant the effects of other observed variables. In policy analysis and in litigation there are many situations in which an expert will put forward a regression model and argue that the estimated regression coefficient associated with the variable of interest can be interpreted as an estimate of the causal effect of that variable, that is, the average change in the outcome that would occur if one were able to change the value of the variable of interest without changing *anything else*. This interpretation requires that all the possible factors that affect the outcome are either (1) included as covariates in the model, or (2) known to be uncorrelated with the variable of interest, so their omission from the model does not lead to bias in the estimated coefficient of the variable of interest.

Causal research designs are techniques to isolate the causal effect of a variable of interest in situations where some determinants of the outcome are omitted, and likely to be correlated with the variable of interest (a problem of omitted variables); or where the variable of interest is itself partly determined by the same unobserved factors that also affect the outcome (a problem of endogeneity).⁸¹ In either case the essential idea is to isolate some part of the variation in the variable of interest that is arguably uncorrelated with the unobserved factors, and to focus on measuring the effect of that more limited variation on the outcome variable.

An illustrative example will make this framework clear. One researcher considers the question of whether an increase in the share of retail outlets owned by

80. See, e.g., Pindyck & Rubinfeld, *supra* note 23, at 178–98 (discussion of instrumental variables estimation).

81. These two cases both give rise to a potential correlation between the variable of interest and the residual term in a basic regression model.

vertically integrated petroleum companies leads to higher gasoline prices.⁸² Such an analysis could be presented as part of an argument that a proposed sale of gas stations owned by an independent retailer to a national company should be blocked, based on what had happened in earlier cases. Unfortunately, a simple regression model relating average gasoline prices (the outcome variable) to the share of vertically integrated sellers (a variable of interest) will not necessarily yield causally interpretable estimates, since the fraction of gas stations operated by vertically integrated sellers, as opposed to independent “non-branded” sellers, may be correlated with unobserved factors not taken into consideration by the statistical model. The author uses a difference-in-differences (DD) design that uses multiple regression to measure the effect of a change in the presence of non-branded sellers in specific market areas on the change in average gasoline prices in those areas, relative to price changes in other areas where there has been no change in the presence of non-branded sellers. She argues that price trends in the other areas are a good benchmark for how prices in the areas affected by the change would have otherwise evolved. In that case, deviations from this benchmark (which is what are measured in the difference-in-differences design) represent the causal effect of the change.

As another example, authors of another study ask how pretrial detention affects the case outcomes and future employment prospects of defendants.⁸³ As with the previous example, it is an open question as to whether the observed relationship between pretrial detention (the variable of interest) and defendant outcomes has a causal interpretation: defendants who receive more lenient bail terms are presumably different from those who do not. To derive a causal estimate, the authors use an instrumental variables design, in which the share of *other* defendants released prior to their trial by the bail judge is used as an instrumental variable for whether the defendant is detained pretrial or not. Since bail judges are usually randomly assigned, the authors argue that average judgments in other cases reflect judge-specific leniency, rather than features of the defendant or case. As a result, their design isolates differences in pretrial release rates, attributable to the luck of the draw in facing a harsh or lenient bail judge, that can be used to measure causal effects on subsequent case outcomes and employment rates.

The subsections that follow describe a range of methodological approaches that can and have been utilized to make causal inferences.

82. Justine S. Hastings, *Vertical Relationships and Competition in Retail Gasoline Markets: Empirical Evidence from Contract Changes in Southern California*, 94 Am. Econ. Rev. 317 (2004), <https://doi.org/10.1257/000282804322970823>.

83. Will Dobbie, Jacob Goldin, & Crystal S. Yang, *The Effects of Pretrial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges*, 108 Am. Econ. Rev. 201 (2018), <https://doi.org/10.1257/aer.20161503>.

Randomized Controlled Trials

The benchmark for causal designs is a simple randomized controlled trial (RCT, also known as a randomized experiment), as is widely used to evaluate new drugs and is sometimes used to analyze novel social programs such as an unemployment benefit for newly released prisoners.⁸⁴ In the simplest RCT, the analyst randomly divides potential subjects into two groups: the treatment group who receive the intervention (for example, the new drug or the unemployment benefits) and the control group who do not.⁸⁵ Because assignment to the two groups is random, the outcomes of the control group provide a valid basis for inferring what the outcomes for the treatment group would have been but for the effects of the treatment, that is, a valid counterfactual. As a result, any systematic differences between the treatment and control groups can be attributed to the intervention, without concern for the presence of omitted or unmeasured variables. An RCT solves the omitted variables problem by assigning the variable of interest to different subjects in a random way, thereby allowing one to assume that all factors that could influence the outcome are equally likely (or “balanced”) in the treatment and control groups, even if these factors cannot be observed.⁸⁶

A widely used conceptual framework for understanding the precise meaning of causality and the power of an RCT to measure the causal effect of an

84. The 1976 Transitional Aid Research Project was a randomized controlled trial in which some newly released prisoners received time-limited weekly benefits while they were out of work. Peter H. Rossi, Richard A. Berk & Kenneth J. Lenihan, *Money, Work, and Crime: Experimental Evidence* (1980). It followed an earlier randomized trial of a similar benefit program for newly released prisoners, known as the Baltimore Living Insurance for Ex-Prisoners project. Charles D. Mallar & Craig V. D. Thornton, *Transitional Aid for Released Prisoners: Evidence from the Life Experiment*, 13 J. Hum. Resch. 208 (1978). Other applications include a 1990s U.S. Census experiment to find whether asking about Social Security numbers would decrease response rates—it did, finding a “3.4% decline in self-response rates attributable to the question.” *New York v. United States DOC*, 351 F. Supp. 3d 502, 526 (S.D.N.Y. 2019). RCTs have been crucial for FDA approval when marketing certain products, including e-cigarettes. *Wages & White Lion Invs., L.L.C. v. FDA*, 41 F.4th 427 (5th Cir. 2022) (upholding FDA decision as neither arbitrary nor capricious, as plaintiffs did not bring sufficient safety evidence in the form of an RCT or longitudinal study).

85. In some cases there will be more than two treatment groups. For example, in the Minneapolis Domestic Violence Experiment (Sherman and Berk, 1984), police officers responding to misdemeanor domestic assaults were randomly assigned to one of three protocols: arrest the suspected offender; offer advice and mediation to the victim and offender; or send the suspect away for eight hours. Strictly speaking this experiment did not have a control group, but any one of the treatment groups can be compared against the other two. Lawrence W. Sherman & Richard A. Berk, *The Specific Deterrent Effects of Arrest for Domestic Assault*, 49 Am. Socio. Rev. 261 (1984), <https://doi.org/10.2307/2095575>.

86. One federal court assessed RCT as the strongest type of evidence: “Scientific evidence is assessed on a scale of Level I to Level V, with Level I granted the most scientific weight and Level V the least. A randomized controlled trial is an example of Level I evidence, expert opinion is Level V.” *McClellan v. I-Flow Corp.*, 710 F. Supp. 2d 1092, 1107 n.10 (D. Or. 2010) (citation omitted).

intervention was proposed by Donald Rubin in 1974.⁸⁷ In this framework, we hypothesize that each subject has two potential outcomes: one if they were to receive the intervention, another if they did not. For example, patients interested in receiving blood pressure medication over the next year would have one blood pressure reading at the end of the year if they were to take the drug, but a different reading if they did not take the drug. The causal effect of the drug for a given subject is the difference between these two potential outcomes. The Average Treatment Effect (ATE) of the drug is the average of the individual-level causal effects.⁸⁸

The fundamental problem of causal inference is that we can see only one of the potential outcomes—depending on whether a subject received the intervention or not—so we can never directly measure the individual treatment effects. In a study that relies on observed behavior (i.e., an observational design), we see the potential outcomes associated with the intervention for subjects that receive the intervention, and the potential outcomes associated with no intervention for those that do not. We could then compare the mean outcomes for the two groups to assess the effect of the intervention. The problem with this comparison is that the average outcome for the nonintervention group can be a misleading estimate of what would have happened to the intervention group in the absence of the intervention. In the blood pressure medication example, the average blood pressure among subjects who do not take the drug could be higher or lower than the average that would have been observed for the intervention group if they had not taken the drug. This difference is defined as *selection bias*: It represents the expected average difference in potential outcomes associated with no intervention between the group that received the intervention and the group that did not.

The sign (positive or negative) and magnitude of selection bias varies from setting to setting and depends on the outcome as well as how the intervention is assigned to (or chosen by) different subjects. For example, in an analysis of the effect of a new medicine on patient death, selection bias will be negative if the medicine is allocated to the sickest patients (because patients who get the medicine are more likely to die, even without the medicine). In contrast, in an analysis of the effect of a training program on earnings, selection bias can be positive if more promising candidates are selected for the program (since they would earn more even without the training).

In an observational study, the simple difference in outcomes between the group that receives the intervention and the group that does not is the sum of the average treatment effect for those who receive the intervention and the selection bias. (See the Appendix for a mathematical statement of this relationship.) In an

87. Donald B. Rubin, *Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies*, 66 J. Educ. Psych. 688 (1974), <https://doi.org/10.1037/h0037350>.

88. For an explanation in a non-RCT case, see *United States v. Brown*, 299 F. Supp. 3d 976, 999 n.16 (N.D. Ill. 2018) (explaining an ATE in a probability-weighting logistic regression).

RCT, the assignment of the intervention is random, removing any selection bias⁸⁹ and ensuring that on average the potential outcomes should be the same for the groups that receive and do not receive the intervention.

This framework provides a slightly different way of thinking about when a multiple regression model, applied to observational data, will yield a coefficient on an indicator variable for being in a group of interest (for example, being over age 40 in an age discrimination case) that can be interpreted causally. In such a setting, the other control variables in the model must fully eliminate any selection bias. In other words, holding constant the observed covariates in the model, all unobserved differences that affect the outcome variable must be the same, on average, for the subjects in the group of interest and those in the rest of the population. This is a very strong condition that needs to be carefully evaluated.

One approach that economists have used to assess the plausibility of the no-selection-bias assumption is to examine how the estimated coefficient on the variable of interest changes as additional control variables are added to the model. In an RCT setting, the estimated coefficient should only change slightly (if at all) as other controls are added to the model, since the key variable of interest is randomly assigned to subjects, and therefore should be uncorrelated with other potential control variables. But in an observational setting, the variable of interest is often correlated with the characteristics of the subjects. In that case, adding control variables may lead to large changes in the estimated coefficient of interest. Altonji et al. suggested that if adding or subtracting observed covariates from the model has little effect on the coefficient of the variable of interest, then it is more plausible that other unobserved factors will have small effects.⁹⁰ A finding of relative stability in the coefficient of interest provides confidence that the estimated effect is robust to changes in the specific set of control variables included in the model, and it may offer some assurance of robustness to unobserved factors under the assumptions proposed by Altonji et al.

Another approach that is sometimes feasible is to show that the variable of interest has no effect on a “placebo” outcome: one that is determined by the same factors as the outcome of interest but logically cannot be affected by the variable

89. Note that selection bias refers to the difference in expected values of the outcome between the group that received the intervention and the group that did not in the absence of the intervention. Although a well-designed RCT will have zero selection bias, there could be differences between the characteristics of the two groups that arise randomly—especially if the groups are small. See Winston Lin, *Agnostic Notes on Regression Adjustments to Experimental Data: Reexamining Freedman's Critique*, 7 *Annals Applied Stat.* 295–318 (2013) (for a discussion of whether these realized differences in characteristics should be accounted for when measuring the treatment effect in an RCT).

90. Joseph G. Altonji, Todd E. Elder & Christopher R. Taber, *Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools*, 113 *J. Pol. Econ.* 151 (2005), <https://doi.org/10.1086/426036>. Emily Oster presents a related analysis that focuses on how the addition of other controls jointly affects the stability of the coefficient of interest and the explanatory power of the model. Emily Oster, *Unobservable Selection and Coefficient Stability: Theory and Evidence*, 37 *J. Bus. & Econ. Stats.* 187 (2019), <https://doi.org/10.1080/07350015.2016.1227711>.

of interest. For example, Rothstein evaluated a regression model that previous researchers had used to show the effect of fourth-grade teacher characteristics on fourth-grade test scores by estimating a similar model but using the same students' *third-grade* test scores as the outcome.⁹¹ He found that fourth-grade teacher characteristics had a positive effect on third-grade scores, suggesting that in this particular setting there was a positive selection bias in the proposed model. Such “falsification” or “placebo” tests—if they are passed—can increase confidence in a causal interpretation of the model.

In cases where an expert is proposing a causal interpretation of the results from a multiple regression model based on observational data, there are two simple questions that need to be addressed:

1. Is there a compelling justification for the assumption (implicit in a causal interpretation) that all omitted factors that could substantially influence the outcome variable are uncorrelated with the variable(s) of interest?
2. What evidence has the expert assembled to support that reasoning?

Pre/Post Design

While randomized controlled experiments offer a very good benchmark, they are rarely conducted in litigation settings, and may not be feasible.⁹² Social scientists and legal experts often rely on other non-experimental approaches (sometimes referred to as non-experimental research designs) that can potentially isolate causal effects by trying to mimic the features of an RCT. The simplest of such approaches, widely used in business practice and policy analysis, is a pre/post design (or before/after design). This approach can be applied in situations where data are available on an outcome for subjects that receive an intervention or treatment both *before* and *after* the intervention. A pre/post design interprets the change in the average outcome across subjects as an estimate of the causal effect of the intervention. For example, an analyst may have data on the number of property crimes in a city in the year before and after a new policing policy is introduced. A pre/post estimate of the effect of the policy is just the change in the number of crimes from the year before to the year after the change.

In a Rubin (1974) framework, a pre/post design uses the average outcome in the pre period as an estimate of the mean potential outcome in the absence of treatment. Since the average outcome in the post period is measured for the same subjects, but with treatment, it may appear that a pre/post difference is free of

91. Jesse Rothstein, *Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement*, 125 Q.J. Econ. 175 (2010), <https://doi.org/10.1162/qjec.2010.125.1.175>.

92. Although in principle an RCT eliminates selection bias, actual RCTs can have other problems (e.g., missing data) that reintroduce selection bias.

selection bias. By measuring the outcomes in different time periods, however, a new source of bias may be introduced, arising from potential changes over time in what would have happened to the average outcomes of the treated subjects, even if they did not receive the intervention. Validity of a pre/post comparison requires that the average outcome for the group that received the treatment would have remained constant in the absence of treatment (i.e., a constant mean assumption).

Many arguments in law, economics, and politics scholarship revolve around interpretations of pre/post designs and the plausibility of the constant mean assumption. For example, Tracey Meares discusses the debate around the interpretation of crime trends in New York City that were used by the city to defend its “Stop, Question, and Frisk” policies in *Floyd v. City of New York*.⁹³ A particular concern that arises is the extent to which the mean outcome among the subjects included in the sample being studied has been driven by the timing of police interventions. As another example, Orley Ashenfelter has noted that many people enroll in government training programs after a job loss.⁹⁴ Even without training, many of these people would be expected to return to work and experience a rebound in earnings.⁹⁵ In such cases, a simple pre/post comparison of earnings for participants in the program will overstate the causal impact of the intervention.

Nevertheless, there are some settings where a pre/post design is compelling, particularly if periods of data are available when there was no change in the treatment. Such data allow an analyst to measure average outcomes for the treated group in the periods before and after the intervention and examine patterns in

93. Tracey L. Meares, *The Law and Social Science of Stop and Frisk*, 10 Ann. Rev. L. & Soc. Sci. 335 (2014), <https://doi.org/10.1146/annurev-lawsocsci-102612-134043>; *Floyd v. City of New York*, 959 F. Supp. 2d 540 (S.D.N.Y. 2013).

94. Orley Ashenfelter, *Estimating the Effect of Training Programs on Earnings*, 60 Rev. Econ. & Stats. 47 (1978), <https://doi.org/10.2307/1924332>. The phenomenon that an intervention tends to be implemented when the mean outcome is trending downward is so common that it is generally referred to as an *Ashenfelter dip*. Ashenfelter dips have been noted in the timing of turnover of professional sports coaches (e.g., Maria De Paola & Vincenzo Scoppa, *The Effects of Managerial Turnover: Evidence from Coach Dismissals in Italian Soccer Teams*, 13 J. Sports Econ. 152 (2012), <https://doi.org/10.1177/1527002511402155>) and school principals (e.g., Ashley Miller, *Principal Turnover and Student Achievement*, 36 Econ. Educ. Rev. 60 (2013), <https://doi.org/10.1016/j.econedurev.2013.05.004>).

95. This is a version of the well-known phenomenon of mean reversion or regression to the mean: the tendency of a statistical process (like individual earnings) to return to its mean value after interruptions that cause unusually high or low values. The concept often arises in securities litigation. See, e.g., *IQ Holdings, Inc. v. Am. Com. Lines Inc.*, No. 6369-VCL, 2013 Del. Ch. LEXIS 234, at *10–12 (Del. Ch. Mar. 18, 2013) (discussing the specific application of mean reversion to calculating the weighted average cost of capital). For another application, in the debate over affirmative action in higher education, see *Students for Fair Admissions, Inc. v. Univ. of N.C.*, 567 F. Supp. 3d 580, 624–25 (M.D.N.C. 2021) (characterizing an expert’s averaging as a “regression to the mean”). For a discussion in the context of a pain control study, see *FTC v. QT, Inc.*, 448 F. Supp. 2d 908, 939 (N.D. Ill. 2006).

these averages over time. A simple plot of these means is known as an event study graph.

In cases where an expert is proposing a causal interpretation of the results from a pre/post design (or as unbiased estimates of the effects represented in a theoretical model), a straightforward question needs to be answered: What evidence has the expert assembled to verify the key pre/post design assumption that absent the intervention, the mean outcomes of subjects affected by the intervention would not have changed?

Difference-in-Differences Design

The difference-in-differences (DD) design (sometimes called *diff-in-diff*) extends the pre/post design by adding a comparison group that is not affected by the intervention or treatment.⁹⁶ In the simplest version of DD, one group of subjects experiences an intervention or treatment at some point in time, while another group does not. For example, David Card and Alan Krueger considered the changes in employment between February and November 1992 at fast food restaurants in New Jersey, where the state minimum wage was increased in April 1992, relative to the changes over the same period at restaurants in Eastern Pennsylvania, where the state minimum wage remained unchanged.⁹⁷ Because the restaurants in Philadelphia and its suburbs were physically close to the restaurants in Camden, New Jersey, and its suburb, it was reasonable to believe that the effects of any other employment-related control covariates would be the same in both locations. In this framework, the DD estimate of the causal effect of the intervention is the change in the average outcome of the treated group, minus the change in the average outcome of the comparison group measured over the same period. This method is widely used in the social sciences and in litigation to examine the causal effects of law changes, firm mergers, and other phenomena.⁹⁸

In the Rubin framework,⁹⁹ a DD design uses the average outcome for the treatment group in the pre period, adjusted by the change in outcomes for the

96. In some studies, the untreated group is called a control group, but that nomenclature risks mixing up control groups in randomized trials and comparison groups in difference-in-difference designs, which are not randomly assigned.

97. David Card & Alan B. Krueger, *Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania*, 84 Am. Econ. Rev. 772 (1994).

98. For example, Hastings (2004) conducts an analysis of retail gasoline prices at two groups of gas stations in Southern California: one group of stations within a mile of a station operated by Thrifty, an independent “non-branded” retailer, and another group of stations further away from any Thrifty station. In the middle of her sample period the Thrifty stations were acquired by ARCO, a major petroleum company; she shows that this led to a rise in prices at stations close to the former Thrifty stations.

99. Rubin, *supra* note 87.

comparison group between the pre and post periods, as an estimate of the counterfactual mean potential outcome for the treatment group in the post period (i.e., as an estimate of the mean of the potential outcome in the absence of treatment for the treatment group in the post period). The key assumption is that in the absence of the intervention the mean potential outcomes for the two groups would move in parallel—a counterfactual assumption that is sometimes called a parallel trends assumption. The DD design assumes parallel trends for the two groups, estimates the change in outcomes that would have occurred in the absence of treatment from the average pre/post change for the comparison group, and then subtracts this value from the average pre/post change for the treatment group to derive a causal estimate.

Like the simpler pre/post design, the potential validity of a DD design can be evaluated by using data on outcomes for the treatment and comparison groups in periods before the intervention. Under the parallel trends assumption, the means of the outcomes for the two groups should move in parallel in the pre-intervention periods. A typical event study graph for a DD design plots the mean outcomes of the treatment and comparison groups in periods before and after the intervention relative to the difference that existed in the period just before the intervention, allowing the analyst to directly assess whether the mean outcomes of the two groups moved in parallel in periods prior to the intervention, and whether the gap between the mean outcomes of the treatment and comparison groups widened or narrowed after the date of the intervention.¹⁰⁰

Generalized Difference-in-Differences Designs

There are many extensions and generalizations of the basic DD design. As with a basic difference-in-differences design, the potential validity of a generalized event-study design can be partly evaluated by plotting the differences in mean outcomes between the treatment and comparison groups for several periods before and after the intervention, removing all differences from the gap in the period immediately before the intervention (i.e., constructing the difference-in-differences between the treatment and comparison groups in each period, relative to the period just before the intervention).¹⁰¹ In a valid design, all the

100. If it is assumed that the intervention causes a once and for all shift in outcomes, then data for the post-intervention periods should also exhibit parallel trends. For an example relating to gasoline stations, see Hastings, *supra* note 82.

101. For example, the method is “commonly accepted” for monetary damage calculations. *Windstream Holdings, Inc. v. Charter Commc'ns Inc.* (*In re Windstream Holdings, Inc.*), 627 B.R. 32, 52–53 (Bankr. S.D.N.Y. 2021). See generally *Messner v. Northshore Univ. HealthSystem*, 669 F.3d 802, 818 (7th Cir.), *reh'g denied*, 2012 U.S. App. LEXIS 4778 (7th Cir. Feb. 28, 2012); *Smith v. Keurig Green Mt., Inc.*, 2020 U.S. Dist. LEXIS 172826, at *26–30 (N.D. Cal. Sept. 21, 2020); *Lowe's Foods LLC v. Burroughs & Chapin Co.*, 2019 U.S. Dist. LEXIS 100410, at *7

difference-in-differences in the pre-intervention period should be close to zero (apart from sampling error) with no discernible trend.

For example, McCrary studies the effect of class-action lawsuits against discriminatory police-hiring policies on the gap between the fraction of African-American officers in a city's police force and the fraction of African-American residents in the city (a difference he calls the "representation gap").¹⁰² He presents graphs that show the difference-in-differences in the representation gap in cities with a lawsuit relative to the year the lawsuit was filed in that city (i.e., treating the filing date of the lawsuit as the date of the intervention). The comparison group includes data from a relatively large number of other cities that had no lawsuit, as well as from cities with lawsuits in future or past years. McCrary's graphs show that prior to the filing date, the representation gap was negative on average (i.e., the African-American share of police officers was lower than the African-American share of city residents) and was stable or even widening slightly prior to a lawsuit, whereas after the filing date the representation gap narrowed, reflecting increases in the relative hiring of African-American officers.

In cases where an expert is proposing a causal interpretation of the results from a difference-in-differences design (i.e., as unbiased estimates of the causal effects of an intervention), two key questions need to be answered:

1. Has the expert assembled evidence on the validity of the parallel trends assumption—that, in the absence of the intervention, the mean outcomes of the treatment group and the comparison group would move in parallel from period to period? Typically, this evidence will take the form of an analysis of data from several periods before the intervention.
2. If such evidence is not available, what other justification or rationale can be offered for the validity of the parallel trends assumption?

Instrumental Variables Design

An instrumental variables (IV) design isolates a specific part of the variation in a variable of interest across subjects and estimates the causal effect of that part of the variation on an outcome. The method relies on the existence of a so-called instrumental variable that satisfies three critical assumptions: (1) it substantially affects the variable of interest; (2) it is uncorrelated with the unobserved

(D.S.C. Apr. 17, 2019); *In re Apple Inc. Device Performance Litig.*, No. 5:18-MD-02827-EJD, 2021 U.S. WL 1022866, (N.D. Cal. Mar. 17, 2021); *Ideker Farms, Inc. v. United States*, 151 Fed. Cl. 560 (2020).

102. Justin McCrary, *The Effect of Court-Ordered Hiring Quotas on the Composition and Quality of Police*, 97 Am. Econ. Rev. 318 (2007), <https://doi.org/10.1257/aer.97.1.318>.

determinants of the outcome; and (3) it has no direct effect on the outcome.¹⁰³ For example, Angrist (1990) addresses the question of whether men who served in the military during the Vietnam era had lower or higher earnings later in life as a result of their service.¹⁰⁴ Given that only a small fraction of men of draft-eligible age actually served in the military in this era, he was concerned about selection biases in the observed differences in earnings between veterans and nonveterans. He therefore used the assignment of a relatively low draft lottery number during the draft lotteries of 1970–1972 as an instrumental variable for serving in the military.

Because of the draft process, Angrist noted that men with low lottery numbers had higher rates of military service than other men. As a result, his proposed instrumental variable satisfies condition (1) above. But since lottery numbers were randomly assigned to different birthdays, it is plausible that having a low lottery number was uncorrelated with all other factors that affect earnings (like family background and cognitive ability), and had no direct effect on earnings, satisfying conditions (2) and (3).

IV designs are often used when there is concern that the main variable of interest is partially determined by the same unobserved factors that also affect the outcome—the situation of an endogenous covariate.¹⁰⁵ In the case of the effect of military service on earnings, for example, one might be concerned that latent health conditions affect the probability of military service and have some effect on earnings. The idea of an IV design is to separate out the part of the endogenous covariate that is predicted by the instrumental variable and use only that part to measure the causal effect. In principle, there may be multiple instrumental variables available, though this can lead to complex issues of

103. IV designs appear prominently in federal antitrust litigation. See *United States v. Aetna Inc.*, 240 F. Supp. 3d 1, 36 (D.D.C. 2017); *In re Domestic Drywall Antitrust Litig.*, 322 F.R.D. 188, 219 (E.D. Pa. 2017); *United States v. Am. Express Co.*, No. 10-CV-4496 (NGG) (RER), 2014 U.S. Dist. LEXIS 87360, at *17–23 (E.D.N.Y. June 24, 2014). See another application in a claim alleging hiring of unauthorized immigrants to depress wages, *Hall v. Thomas*, 753 F. Supp. 2d 1113, 1145–47 (N.D. Ala. 2010).

104. Joshua D. Angrist, *Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records*, 80 Am. Econ. Rev. 313 (1990). In a similar vein, a large number of studies—summarized in Julia Chabrier, Sarah Cohodes & Philip Oreopoulos, *What Can We Learn from Charter School Lotteries?*, 30 J. Econ. Persps. 57 (2016), <https://doi.org/10.1257/jep.30.3.57>—use the outcomes of admissions lotteries as instrumental variables for attending a charter school in analyzing the effects of charter schools on student test scores.

105. For example, schooling is often modeled as an endogenous regressor in models that relate earnings to schooling, reflecting the belief that some of the same unobserved factors that determine schooling—like ability, ambition, and parental encouragement—also partly determine earnings. David Card, *The Causal Effect of Education on Earnings*, in 3 *Handbook of Labor Economics* (Orley Ashenfelter & David Card eds., 1999).

interpretation and proper inference.¹⁰⁶ This presentation focuses on the simpler case of a single instrumental variable (also known as the just-identified case).

The IV method with a single instrumental variable involves estimating two regression models: hence the term *two-stage least squares* that is widely used to describe the method. The first of these (known as the *first-stage model*) regresses the endogenous variable on the instrumental variable (and any other control variables included in the model). The estimates from this model are then used (in the second stage) to form predicted values of the endogenous regressor for each unit (or subject) in the dataset. Since the prediction is just a weighted average of the instrumental variable and the other control variables, and by assumption the instrumental variable is uncorrelated with the unobserved determinants of the outcome, the associated predictions are also uncorrelated with the unobserved factor. The second-stage model is then a multivariate regression model that relates the outcome to the predicted values of the endogenous covariate and the same control variables included in the first-stage model. The estimated coefficient for the predicted variable is interpreted as a causal estimate of the effect of the endogenous covariate on the outcome.

The key coefficient of the predicted endogenous covariate in the second-stage model can be derived in a different way that illustrates its interpretation. Consider a third model, known as the reduced-form model, that regresses the outcome variable on the instrumental variable and the other controls.¹⁰⁷ The coefficient of the instrumental variable in this model can be interpreted causally because it is assumed to be uncorrelated with the unobserved determinants of the outcome. This coefficient expresses the expected change in the outcome variable for a unit change in the value of the instrumental variable. But since the instrumental variable is assumed to have no *direct* effect on the output, this reduced-form result must arise via the effect of the instrumental variable on the endogenous covariate.

Consider, for example, the analysis of the effect of military service on earnings, using a low lottery number as the instrumental variable. Angrist (in his Table 3) shows that having a low lottery number is associated with about \$400 lower earnings per year for men born in 1950 than for men with higher lottery numbers. This is the reduced-form effect of a low lottery number. The entirety of this difference is presumably attributable to the fact that these men were more likely to serve in the military. In fact, Angrist shows that the first-stage model relating military service to lottery numbers implies that they had a

106. John Bound, David A. Jaeger & Regina M. Baker, *Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable Is Weak*, 90 J. Am. Stat. Ass'n 443 (1995), <https://doi.org/10.2307/2291055>; James Stock, Jonathan Wright & Mohiro Yugo, *A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments*, 20 J. Bus. & Econ. Stats. 518 (2002), <https://doi.org/10.1198/073500102288618658>.

107. The set of variables included in the reduced-form model must be the same as the set included in the first-stage model.

16 percentage-point higher probability of serving in the military than men with higher lottery numbers. Using mathematical notation, suppose that serving in the military has a causal effect of β_1 dollars.. Then the reduced-form effect ($-\$400$) must arise because a 0.16 increase in the probability of military service, multiplied by the effect of military service on earnings (β_1) is $-\$400$. In other words, $-400 = \beta_1 \times 0.16$, implying that $\beta_1 = -400/0.16 = -\$2500$.

In general, the reduced-form effect of the instrumental variable on the outcome is the product of the first-stage effect of the instrumental variable on the endogenous covariate, and the causal effect of a change in the endogenous covariate on the output. This means that one can unscramble the causal effect by dividing the reduced-form coefficient by the first-stage coefficient. (See the Appendix for a mathematical statement of this procedure.) In the case where there is a single instrumental variable, this ratio is exactly equal to the estimated coefficient of the predicted endogenous covariate in the second-stage model.

As another example of the use of an IV design, Dobbie et al. relate defendant outcomes in criminal cases—including case outcomes such as the incidence of guilty pleas and longer-term outcomes like employment rates—to an indicator variable of whether the defendant was detained prior to their trial.¹⁰⁸ Although the authors' dataset includes relatively detailed controls for case characteristics, the authors are concerned that judges set bail based in part on *other* case characteristics that are not in the dataset, and that might also affect later outcomes for the defendant. In essence, pretrial detention is an endogenous covariate.

As an instrumental variable for pretrial detention, the authors use the average rate of pretrial detention in other cases handled by the same bail judge.¹⁰⁹ The plausibility of this choice rests on the fact that in the two large jurisdictions in their sample, bail judges are typically assigned at random to cases (conditional on time of day and day of the week of the hearing). The authors show that judges with a higher rate of pretrial detention in other cases are more likely to set harsher bail terms that lead to a higher probability of detention, satisfying condition (1) above. They also show that the instrumental variable is uncorrelated with a long list of characteristics that are highly predictive of pretrial detention (such as measures of the defendant's previous crime record). This lack of correlation does not *prove* that conditions (2) and (3) are satisfied but is consistent with their assertion that judges are as good as randomly assigned and would fail if harsher bail judges were assigned to specific types of cases. The authors' analysis of the judge assignment process

108. Dobbie et al., *supra* note 83.

109. This is known as the leave-out mean. To account for the fact that certain kinds of cases are more likely at certain times or days of the week, Dobbie et al. use a leave-out mean of a regression-adjusted pretrial detention rate, where the adjustment factors include time-of-day and day-of-week indicators. *Id.*

makes a plausible case that conditions (2) and (3) are satisfied.¹¹⁰ Of course, not all assignments are random, in part because not all judges are on the same selection “wheel.” However, empirical evidence supports the view that for broad categories of cases, the judicial assignment process appears to be random.¹¹¹

As in the Rubin framework, each subject has two potential outcomes, representing the value of their outcome if they receive the intervention or not.¹¹² Among all the subjects assigned the high value of the instrumental variable, there is a fraction of compliers whose observed outcome is their potential outcome with treatment. Among the subjects assigned the low value of the instrumental variable are the same fraction of compliers whose observed outcomes are their potential outcomes without treatment. The share of compliers in the two groups, and their distributions of potential outcomes, is the same in both cases because the instrumental variable is working as if the individuals were randomly assigned. As a result, the expected difference in average outcomes between subjects that are assigned high and low values of the instrumental variable is the difference in potential outcomes among the compliers, multiplied by their share in the sample. This product is estimated by the coefficient of the instrumental variable in the reduced-form equation of the IV procedure. Dividing this coefficient by the first-stage coefficient of the instrumental variable (which, as noted earlier, is the IV estimate of the causal effect) yields an estimate of the average treatment effect among the compliers. In many applications the complier group can be relatively small: For example, in Angrist’s study of the Vietnam-era draft lottery, the group was composed of only about 15% of draft-age men.¹¹³ Thus the estimated effect derived from an IV design does not necessarily provide an estimate of the average treatment effect for the entire subject population.

110. For an extension of Rubin’s potential-outcomes framework to an IV design in which the endogenous covariate is an indicator for receiving an intervention (e.g., serve in the military or not), and the instrumental variable is also dichotomous (e.g., receive a high or low draft lottery number) that is distributed randomly or (by assumption) *as if* randomly, see Guido W. Imbens & Joshua D. Angrist, *Identification and Estimation of Local Average Treatment Effects*, 62 *Econometrica* 467 (1994), <https://doi.org/10.2307/2951620>. For example, an instrumental variable based on the last four digits of a Social Security number is not randomly assigned but may be “as good as” randomly assigned. Often it is assumed that a variable is as good as randomly assigned conditional on a set of basic controls. For example, whether a person wins a lottery is random holding constant the number of tickets they held, but not unconditionally.

111. Orley Ashenfelter, Theodore Eisenberg & Stewart J. Schwab, *Politics and the Judiciary: The Influence of Judicial Background on Case Outcomes*, 24 *J. Legal Stud.* 257 (1995), <https://doi.org/10.1086/467960>.

112. The never takers and always takers also have two potential outcomes, but their outcomes are not affected by which value of the instrumental variable they are assigned, because the instrumental variable does not affect their participation in the intervention, and by assumption changes in the instrumental variable do not directly affect outcomes.

113. Angrist, *supra* note 104.

In applications of IV it is important to assess the validity of the three assumptions noted above. Violations of these assumptions can lead to large biases in the estimated causal effects obtained from an IV design—in some cases, even larger than the bias associated with a simple multivariate regression model that ignores concerns about the endogenous variable (i.e., the purely observational design, estimated by ordinary least squares (OLS)). With that in mind, an IV estimate should always be compared directly with the associated OLS estimate. If the two estimates are substantively different, there should also be a careful discussion of the sources of bias in the OLS estimate.

The first assumption for an IV design is that the instrumental variable has a substantial effect on the variable of interest. Since this effect is measured by the coefficient of the instrumental variable in the first-stage model, standard practice is to present the first-stage model and evaluate the sign (positive or negative), size, and statistical significance of this coefficient. The sign is important because the logic of an IV design rests on the idea that changes in the instrumental variable cause systematic changes in the endogenous covariate: the direction of this causal effect is reflected in the sign of the first-stage coefficient and must be interpretable.¹¹⁴ Size and significance are important because the IV estimate of the causal effect is based on the response of the outcome to the part of the endogenous covariate that is attributable to variation in the instrumental variable. When the first-stage coefficient is large in magnitude and highly statistically significant, this part of the variation in the endogenous covariate can be more reliably identified.¹¹⁵ As an example of how sign, size, and significance can be assessed visually, Dobbie et al. plot the mean of their endogenous covariate for various ranges of their instrumental variable and show that there is a strong relationship of the expected sign.¹¹⁶ In addition, they report the estimated coefficient of the instrumental variable in their first-stage equation.

The second assumption is that the instrumental variable is uncorrelated with all unobserved determinants of the outcome. This condition will be satisfied if

114. Isaiah Andrews and Timothy B. Armstrong note that use of information on the sign of the first-stage coefficient can substantially improve the performance of IV estimators. Isaiah Andrews & Timothy B. Armstrong, *Unbiased Instrumental Variables Estimation Under Known First-Stage Sign*, 8 Quantitative Economics, 479 (2017), <https://doi.org/10.3982/QE700>.

115. In the case of multiple instrumental variables, a conventional “rule of thumb” is that the F-statistic that provides a means of testing the goodness of fit of the instrumental variables in the first-stage equation has to exceed ten. James Stock, Jonathan Wright & Mohiro Yugo, *A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments*, 20 J. Bus. & Econ. Stats. 518 (2002), <https://doi.org/10.1198/073500102288618658>. With a single instrumental variable, such a cutoff is not necessarily best practice. Instead, it is recommended that analysts report so-called Anderson-Rubin confidence intervals for the estimated causal effect. See Isaiah Andrews, James H. Stock & Liyang Sun, *Weak Instruments in IV Regression: Theory and Practice*, 11 Ann. Rev. Econ. 727 (2019), <https://doi.org/10.1146/annurev-economics-080218-025643>. These confidence intervals incorporate uncertainty from the first-stage model.

116. Dobbie et al., *supra* note 83.

the instrumental variable is randomly assigned. Otherwise, the burden of persuasion should be on the analyst to present evidence that the instrumental variable is as good as randomly assigned, holding constant a basic set of control variables in the model. One approach (used by Dobbie et al.) is to show that the instrumental variable is uncorrelated with predetermined subject characteristics that are predictive of the outcome (at least once the main control variables are introduced).¹¹⁷ Another complementary approach is to show that the reduced-form effect of the instrumental variable on the outcome remains relatively stable as additional control variables are added to the model (holding constant the set of basic controls needed to ensure the instrumental variable is as good as randomly assigned). This is an adaptation of the argument of Altonji et al. for observational designs previously discussed.¹¹⁸ But since the reduced-form regression is itself just a multiple regression model with one key covariate of interest (the instrumental variable), the same arguments apply.

The third assumption is that the instrumental variable does not itself directly affect the outcome—sometimes called an exclusion restriction. There are two related concerns over this assumption in particular settings. First, sometimes an instrumental variable may in fact have its own causal effect.¹¹⁹ This is not normally an issue if the instrumental variable is based on a randomizing process, as in the draft-lottery study of Angrist,¹²⁰ or a quasi-randomizing process, as in the study based on judge assignments by Dobbie et al.¹²¹ In other settings, however, the exclusion restriction must be justified on theoretical grounds. For example, in models of consumer demand for differentiated products that are widely used in antitrust analyses, some analysts argue that sums of characteristics of competing products, interacted with whether the competing products are supplied by the same firm or another firm, are plausible instrumental variables for product prices.

Second, even if the instrumental variable is as good as randomly assigned, there may be another endogenous determinant of the outcome that is also affected by the instrumental variable. In this case, the IV design confounds the effects of the two endogenous variables—a so-called multiple channel problem—possibly overstating or understating the causal effect of the variable of interest. For

117. Such an analysis is conducted by Chan et al., who show that medical patient characteristics help predict the probability of a pneumonia diagnosis (their outcome variable) but do not predict the instrumental variable in their model once they control for patient age, a variable they argue is needed to ensure that the instrumental variable's distribution is as good as random across patients. David Chan, Matthew Gentzkow & Chuan Yu, *Selection with Variation in Diagnostic Skill: Evidence from Radiologists*, 137 Q.J. Econ. 729 (2022), <https://doi.org/10.1093/qje/qjab048>.

118. Altonji et al., *supra* note 90.

119. For example, some studies of the effect of education on earnings have used parental education as an instrumental variable. See Card, *supra* note 105.

120. Angrist, *supra* note 104.

121. Dobbie et al., *supra* note 83.

example, an instrumental variable that increases the education level of young people of a given age also typically reduces the years of work experience they have completed since finishing their schooling (since someone of a given age with more education has had less time to work). Since both education and work experience affect earnings, it may be inappropriate to attribute the difference in earnings between people with different values of the instrumental variable entirely to their differences in schooling. Arguments ruling out such multiple-channel concerns are typically based on theoretical grounds.

In cases where an expert is proposing a causal interpretation of the results from an instrumental variables design (or as unbiased estimates of the effects represented in a theoretical model), a series of key questions must be answered:

1. What is the instrumental variable (or set of instrumental variables)? What is the reasoning underlying the use of this instrumental variable?
2. What is the sign (positive or negative) and size of the coefficient representing the effect of the instrumental variable in the first-stage model? The method depends critically on the assumption that the first-stage model is describing the effect of the instrumental variable on the variable of interest. Therefore, the first-stage equation must be interpretable; it must make sense in light of the claim at issue.
3. Is the measured effect of the instrumental variable in the first-stage model statistically significant?
4. What is the reasoning underlying the implicit assumption—necessary for a causal interpretation of estimates from an instrumental variables design—that the instrumental variable is uncorrelated with any unobserved determinants of the outcome variable in the second-stage model?
5. What evidence has the expert assembled to support that reasoning?
6. What is the reasoning underlying the implicit assumption—necessary for a causal interpretation of estimates from an instrumental variables design—that the instrumental variable has no direct effect on the outcome variable?

More Advanced Regression Methods

Fixed-Effects and Random-Effects Regression Models

As the volume of data that is available has grown over time, social scientists have often utilized both time-series and cross-section regression models that allow for the possibility that differences in the values of the dependent variable can be captured in differences in the constant term in the regression. A fixed-effects regression model will include a series of dummy variables as well as the usual set of

relevant covariates.¹²² In a cross-section regression model, the fixed effects might account for variation in certain characteristics of individuals in the sample, such as geographic location.¹²³ In a time-series model that includes daily or weekly data, the fixed-effects model might account for certain time periods, such as year or month.

There are benefits and costs associated with the inclusion of a group of fixed-effects variables in a regression model. The benefit is that the fixed-effects model can reduce or eliminate the possibility of omitted variable bias. Suppose, for example, that in a study of the effect of a merger on the prices paid by consumers for different models and brands of clothes dryers, one believes it appropriate to include a set of covariates that account for the brand, the size of the machine, the number of cycles, and whether the dryer is electric or gas.¹²⁴ As an alternative, however, the inclusion of a series of fixed-effects dummy variables representing each individual brand and model of clothes dryer might provide a richer analysis.

The cost is that the inclusion of the fixed-effects variables might mask a more complete characterization of the phenomenon being studied. To continue the clothes dryer example, suppose that companies frequently introduce (and withdraw) new models, with each model typically only lasting in the market a year or less. In this case, controlling for individual model effects may preclude a complete analysis of the effect of the merger on prices, since only models that were in the market both before and after the merger contribute to the estimation of the effect of the merger on prices. A more instructive regression model would have a more limited set of controls that allow pre- and post-merger comparisons to be made across all the models offered in the market in each month, rather than just the subset of brands that were present in the market before and after the merger.

The fixed-effects model offers a reasonable approach when we believe that the differences in the dependent variable measured across units of observation can be viewed as vertical shifts in the regression line, and when the basic units of interest (in the example, models of clothes dryers) are observed across all time periods. But if the classification of basic units varies over time (as in the dryer example), it may be more appropriate to consider a model with a set of covariates that are observed consistently, and to assume that the variation in the outcome variable at the group level is an extra error component that is randomly

122. See, e.g., *Conrad v. Jimmy John's Franchise, LLC*, No. 18-CV-00133, 2021 U.S. Dist. LEXIS 33933 (N.D. Ill. Jan. 10, 2022); *United States v. Mariner Health Care*, 552 F. Supp. 3d 938 (N.D. Cal. 2021); *Sidibe v. Sutter Health*, No. 12-CV-04854, 2020 U.S. Dist. LEXIS 136657 (N.D. Cal. July 30, 2020); *Nexteel Co. v. United States*, 569 F. Supp. 3d 1354 (Ct. Int'l Trade 2022); *In re Allstate Corp. Sec. Litig.*, No. 16-C-10510, 2022 U.S. Dist. LEXIS 52616 (N.D. Ill. 2022).

123. See, e.g., *Annex Books, Inc. v. City of Indianapolis*, 581 F.3d 460 (7th Cir. 2009).

124. See Orley C. Ashenfelter, Daniel S. Hosken & Matthew C. Weinberg, *The Price Effects of a Large Merger of Manufacturers: A Case Study of Maytag-Whirlpool*, 5 Econ. J.: Econ. Pol'y 239 (2013), <https://doi.org/10.1257/pol.5.1.239>.

distributed across the individual units of observation.¹²⁵ In essence, using a random-effects model is more appropriate when the goal of a study is to understand behavior in the underlying population of all possible cross-sectional units, whereas the fixed-effects model is focused on an understanding of the behavior of units within the groups that are identified by the fixed effects.

Methods for Estimating Regression Models

The estimated parameters of the basic linear multiple regression model (i.e., the regression coefficients) are most easily and most often estimated using computer programs that seek to minimize the sum of the squares of the regression residuals. When the errors in the regression model fit a normal distribution, the parameter estimates will lie within a normal distribution whose mean is an unbiased estimate of the population mean and whose variance is the minimum variance among all possible estimators. When certain assumptions of the model are violated, the least-squares method can still be highly effective. For example, when heteroscedasticity is an issue (the error variance itself is known to be variable), then weighted least squares can be appropriate. In this case, the data might be weighted by the inverse of an estimate of the standard deviation of the error term.

When the errors are believed to be far from normally distributed, maximum likelihood estimation offers a more robust methodology than least squares. For example, probit and logit models that are widely used in studies of binary outcome variables are fit by this method, as are models of censored outcomes (such as the length of time that a patient survives after a medical procedure, which will be censored if the patient is still alive at the time the data are collected). In essence, maximum likelihood estimation determines values for the parameters of the regression model that maximize the likelihood that the process described by the model produced the data that were observed. The methodology provides a flexible approach that is suitable for a wide variety of models, including models for which the basic regression framework is unsuitable.¹²⁶ It has the advantage that, for large samples, the resulting estimates will be unbiased, and if

125. See, e.g., *Newman v. McNeil Consumer Healthcare*, No. 10-C-1541, 2013 U.S. Dist. LEXIS 113438 (N.D. Ill. Mar. 29, 2013); *In re Lipitor (Atorvastatin Calcium) Mktg. Sales Pracs. & Prods. Liab. Litig.*, 145 F. Supp. 3d 573 (D.S.C. 2015).

126. This approach is often used in voting rights cases. See *Robinson v. Ardoin*, 605 F. Supp. 3d 759 (M.D. La. 2022) (describing the ecological inference (EI) model utilized by an expert as using “maximum likelihood statistics to produce estimate of voting patterns by race”) (citation omitted); *Fabela v. City of Farmers Branch*, No. 3:10-CV-1425-D, 2012 U.S. Dist. LEXIS 108086, at *38 n.22 (N.D. Tex. Aug. 2, 2012) (likewise detailing that the benefit of the EI model in analyzing election data is that it does not rely “on an assumption of linearity but instead uses a maximum likelihood estimation” that “provides accurate confidence intervals”).

the model is correctly specified, those estimates will have the smallest possible variance.

Why not use maximum likelihood estimation in all cases? In a basic regression model, if one assumes that the errors are normally distributed, then a standard regression approach yields estimated coefficients that are the same as the maximum likelihood estimates. Thus, in many applications, the two approaches are the same. More generally, a standard-regression approach has the advantage that the effects of problems like omitted variables, endogeneity, and measurement error in the explanatory variables are well understood, and often can be assessed in ways that have been extensively developed in the applied econometrics literature.

Measuring the goodness of fit in a model that is estimated using a maximum likelihood method can be complicated, especially when the model is inherently nonlinear and when there is a complex error structure. In simple probit and logit models, measures of goodness of fit known as pseudo *R*-squared measures are widely used.¹²⁷ Different versions of models fit by maximum likelihood (say, with more or fewer covariates) can also be compared by a likelihood-ratio test, which measures the percentage improvement in the likelihood of the more complex model over the simpler model.¹²⁸

The Expert

Multiple regression and other forms of complex statistical models are taught to students in extremely diverse fields, including but not limited to statistics, economics, political science, sociology, psychology, anthropology, public health, and history. Nonetheless, the methodology is difficult to master, necessitating a combination of technical skills (the science) and experience (the art). This naturally raises two questions:

1. Who should be qualified as an expert?
2. When and how should the court appoint an expert to assist in the evaluation of issues relating to multiple regression?

127. One version of pseudo *R*-squared was developed in Daniel McFadden, *Conditional Logit Analysis of Qualitative Choice Behavior*, in *Frontiers in Econometrics* 105 (Paul Zarembka ed., 1973).

128. The significance of the improvement in the model can be evaluated using the chi-square distribution. The chi-square distribution is a standard reference distribution in statistics that represents the sum of the squares of a group of independent standardized normal random variables.

Who Should Be Qualified as an Expert?

Any individual with substantial training in and experience with multiple regression and other statistical methods may be qualified as an expert. A doctoral degree in a discipline that teaches theoretical or applied statistics, such as economics, history, and psychology, usually signifies to other scientists that the proposed expert meets this preliminary test of the qualification process. It is noteworthy that a proposed expert whose only statistical tool is regression analysis may not be able to judge when a statistical analysis should be based on an approach other than regression analysis.¹²⁹

The decision to qualify an expert in regression analysis rests with the court. Clearly, the proposed expert should be able to demonstrate an understanding of the discipline. Publications relating to regression analysis in peer-reviewed journals, active memberships in related professional organizations, courses taught on regression methods, and practical experience with regression analysis can indicate a professional's expertise. However, the expert's background and experience with the specific issues and tools that are applicable to a particular case should also be considered during the qualification process. Thus, if the regression methods are being utilized to evaluate damages in an antitrust case, the qualified expert should have sufficient qualifications in economic analysis as well as statistics. An individual whose expertise lies solely with statistics will have a limited ability to evaluate the usefulness of alternative economic models. Similarly, if a case involves eyewitness identification, a background in psychology as well as statistics may provide essential qualifying elements.

Should the Court Appoint a Neutral Expert?

There are conflicting views on the issue of whether court-appointed experts should be used. In complex cases in which two experts are presenting conflicting statistical evidence, the use of a "neutral" court-appointed expert can be advantageous, although courts would need to find the funds to appoint such an expert. There are those who believe, however, that there is no such thing as a truly neutral expert. In any event, if an expert is chosen, that individual should have substantial expertise and experience—ideally, the expert should be someone who is respected (and trusted) by both plaintiffs and defendants.¹³⁰

129. To illustrate, a case involving allegations of the pass-through of a manufacturers' price-fixing agreement may require testimony by experts with statistical as well as economics training.

130. Judge Posner notes in *In re High Fructose Corn Syrup Antitrust Litig.*, 295 F.2d 651, 665 (7th Cir. 2002), "the judge and jury can repose a degree of confidence in his testimony that it could not repose in that of a party's witness. The judge and the jury may not understand the neutral expert perfectly but at least they will know that he has no axe to grind, and so, to a degree anyway, they will

The appointment of such an expert is likely to influence the presentation of the statistical evidence by the experts for the parties in the litigation. The neutral expert will have an incentive to present a balanced position that relies on broad principles for which there is consensus amongst the community of experts. As a result, the parties' experts can be expected to present testimony that confronts core issues that are likely to be of concern to the court and testimony that is sufficiently balanced to be persuasive to the court-appointed expert.¹³¹

Rule 706 of the Federal Rules of Evidence governs the selection and instruction of court-appointed experts. In particular,

1. The expert should be notified of the duties through a written court order or at a conference with the parties.
2. The expert should inform the parties of findings orally or in writing.
3. If deemed appropriate by the court, the expert should be available to testify and may be deposed or cross-examined by any party.
4. The court must determine the expert's compensation.¹³²
5. The parties should be free to utilize their own experts.

Although not required by Rule 706, it will usually be advantageous for the court to opt for the appointment of a neutral expert as early in the litigation process as possible. It will also be advantageous to minimize any *ex parte* contact with the neutral expert; this will diminish the possibility that one or both parties will come to the view that the court's ultimate opinion was unreasonably influenced by the neutral expert.

Rule 706 does not offer specifics as to the process of appointment of a court-appointed expert. One possibility is to have the parties offer a short list of possible appointees. If there was no common choice, the court could select from the combined list, perhaps after allowing each party to exercise one or more peremptory challenges. Another possibility is to obtain a list of recommended experts from a selection of individuals known to be experts in the field.

be able to take his testimony on faith." Such an appointment is not an obligation of Federal Rule of Evidence 706. *See Stevenson v. Windmoeller & Hoelscher Corp.*, 39 F.4th 466 (7th Cir. 2022).

131. For a discussion of the presentation of expert evidence generally, including the use of court-appointed experts, see Samuel R. Gross, *Expert Evidence*, 1991 Wis. L. Rev. 1113 (1991). Some critique court-appointed experts as biased for the prosecution. *See* J. Blais & A.E. Forth, *Prosecution-Retained Versus Court-Appointed Experts: Comparing and Contrasting Risk Assessment Reports in Preventative Detention Hearings*, 38 L. & Hum. Behav., 531 (2014), <https://doi.org/10.1037/lhb0000082>. For one study evaluating the newer practice of "hot tubbing," or concurrent expert testimony, see Jennifer T. Perillo et al., *Testing the Waters: An Investigation of the Impact of Hot Tubbing on Experts from Referral Through Testimony*, 45 L. & Hum. Behav. 229 (2021), <https://doi.org/10.1037/lhb0000446>.

132. Although Rule 706 states that the compensation must come from public funds, complex litigation may be sufficiently costly as to require that the parties share the costs of the neutral expert.

Presentation of Statistical Evidence

The costs of evaluating statistical evidence can be reduced and the precision of that evidence increased if the discovery process is used effectively. In evaluating the admissibility of statistical evidence, courts should consider the following issues:

1. Has the expert provided sufficient information to replicate the multiple regression analysis?
2. Are the expert's methodological choices reasonable or are they arbitrary and unjustified?
3. What disagreements exist regarding the data on which the analysis is based?

In general, a clear and comprehensive statement of the underlying research methodology is a requisite part of the discovery process. The expert should be required to reveal both the nature of the experimentation carried out and the sensitivity of the results to the data and to the methodology.

The following suggestions can substantially improve the discovery process:

1. To the extent possible, the parties should be encouraged to agree to use a common database. Even if disagreement about the significance of the data remains, early agreement on a common database can help focus the discovery process on the important issues in the case.
2. A party that offers data to be used in statistical work, including multiple regression analysis, should be encouraged to provide the following to the other parties: (a) a detailed record of the underlying sources of the data; (b) a data file (ideally, in a commonly used format such as comma-separated values format); (c) complete documentation of the variables in the file; (d) computer programs that were used to generate the data; and (e) explanatory notes associated with the computer programs. The documentation should be sufficiently complete and clear so that the opposing expert can reproduce all the statistical work.
3. A party offering data should make available the personnel involved in the compilation of such data to answer the other party's technical questions concerning the data and the methods of collection or compilation.
4. A party proposing to offer an expert's regression analysis at trial should ask the expert to fully disclose, along with the database and its sources,¹³³ (a) the method of collecting the data and (b) the methods of analysis. When possible, this disclosure should be made sufficiently in advance of trial so that the opposing party can consult its experts and prepare

133. These sources would include all variables used in the statistical analyses conducted by the expert, not simply those variables used in a final analysis on which the expert expects to rely.

cross-examination. The court must decide on a case-by-case basis where to draw the disclosure line.

These suggestions are motivated by the objective of improving the discovery process to make it more informative. The fact that these questions may raise some doubts or concerns about a particular regression model should not be taken to mean that the model does not provide useful information. It does, however, take considerable skill for an expert to determine the extent to which information is useful when the model being utilized has some shortcomings.

Which Database Information and Analytical Procedures Will Aid in Resolving Disputes Over Statistical Studies?

To help resolve disputes over statistical studies, an expert should follow the guidelines below when presenting database information and analytical procedures.¹³⁴

The expert should

1. Clearly state the objectives of the study, as well as the time frame to which it applies and the statistical population to which the results are being projected.
2. Report the units of observation (e.g., consumers, businesses, or employees).
3. Clearly define each variable.
4. Clearly identify the sample for which data are being studied,¹³⁵ as well as the method by which the sample was obtained.
5. Reveal if there are missing data, whether caused by a lack of availability (e.g., in business data) or nonresponse (e.g., in survey data), and the method used to handle the missing data (e.g., deletion of observations).
6. Report investigations into errors associated with the choice of variables and assumptions underlying the regression model.
7. If samples were chosen randomly from a population (i.e., probability sampling procedures were used),¹³⁶ make a good-faith effort to provide an estimate of a sampling error, the measure of the difference between the

134. For a more complete discussion of these requirements, see *The Evolving Role of Statistical Assessments as Evidence in the Courts* 256 (Stephen E. Fienberg ed., 1989) (Recommended Standards on Disclosure of Procedures Used for Statistical Studies to Collect Data Submitted in Evidence in Legal Cases).

135. The sample information is important because it allows the expert to make inferences about the underlying population.

136. In probability sampling, each representative of the population has a known probability of being in the sample. Probability sampling is ideal because it is highly structured, and in principle

sample estimate of a parameter (such as the mean of a dependent variable under study), and the (unknown) population parameter (the population mean of the variable).¹³⁷

8. Report the set of procedures that were used to minimize sampling errors if probability sampling procedures were not used.

Appendix: The Basics of Multiple Regression

Introduction

This appendix illustrates, through examples, the basics of multiple regression analysis in legal proceedings. Often, visual displays are used to describe the relationship between variables that are used in multiple regression analysis. Figure 2 is a scatterplot that relates scores on a job aptitude test (shown on the x -axis) and job performance ratings (shown on the y -axis). Each point on the scatterplot shows what a particular individual scored on the job aptitude test and how that individual's job performance was rated. For example, the individual represented by Point A in Figure 2 scored 49 on the job aptitude test and had a job performance rating of 62.

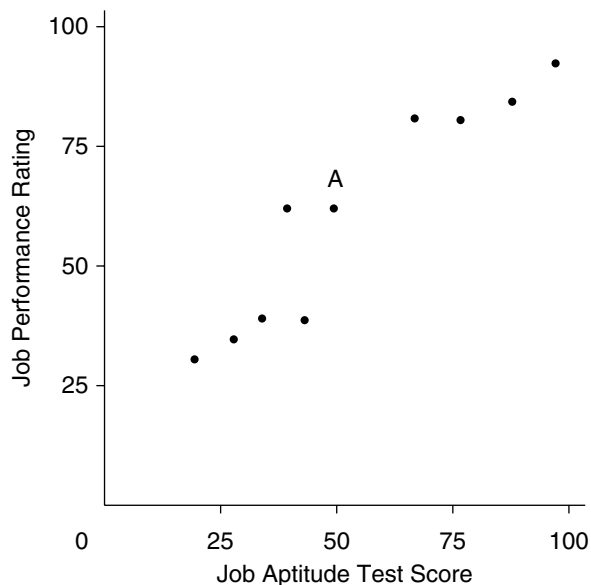
The degree of linear association between two variables can be summarized by a correlation coefficient, which ranges in value from -1 (a perfect linear negative relationship) to $+1$ (a perfect linear positive relationship).¹³⁸ Figure 3 depicts three possible relationships between the job-aptitude variable and the job-performance variable. In Figure 3(a), there is a positive correlation: In general, higher job performance ratings are associated with higher aptitude test scores, and lower job performance ratings are associated with lower aptitude test scores. In Figure 3(b), the correlation is negative, meaning that higher job performance ratings are associated with lower aptitude test scores, and lower job performance ratings are associated with higher aptitude test scores. Positive and negative correlations can

it can be replicated by others. Nonprobability sampling is less desirable because it is often subjective, relying to a large extent on the judgment of the expert.

137. Sampling error is often reported in terms of standard errors or confidence intervals. See Appendix, below, for details.

138. Loosely speaking, the correlation coefficient summarizes the extent to which a scatter plot of variable Y against variable X lies on a line. In fact, the R -squared of a simple linear regression model relating variable Y to variable X and a constant is the square of the correlation coefficient (which is the reason for the name R -squared). In cases where Y and X are related, but in a nonlinear way, the correlation coefficient is a less useful summary. For example, if Y is perfectly determined as a quadratic function of X , and X takes on values of -2 , -1 , 0 , 1 , and 2 , then the correlation coefficient relating to Y and X is precisely 0 .

Figure 2. Scatterplot of scores on a job aptitude test relative to job performance rating.



be relatively strong or relatively weak. If the relationship is sufficiently weak, there is effectively no correlation, as is illustrated in Figure 3(c).

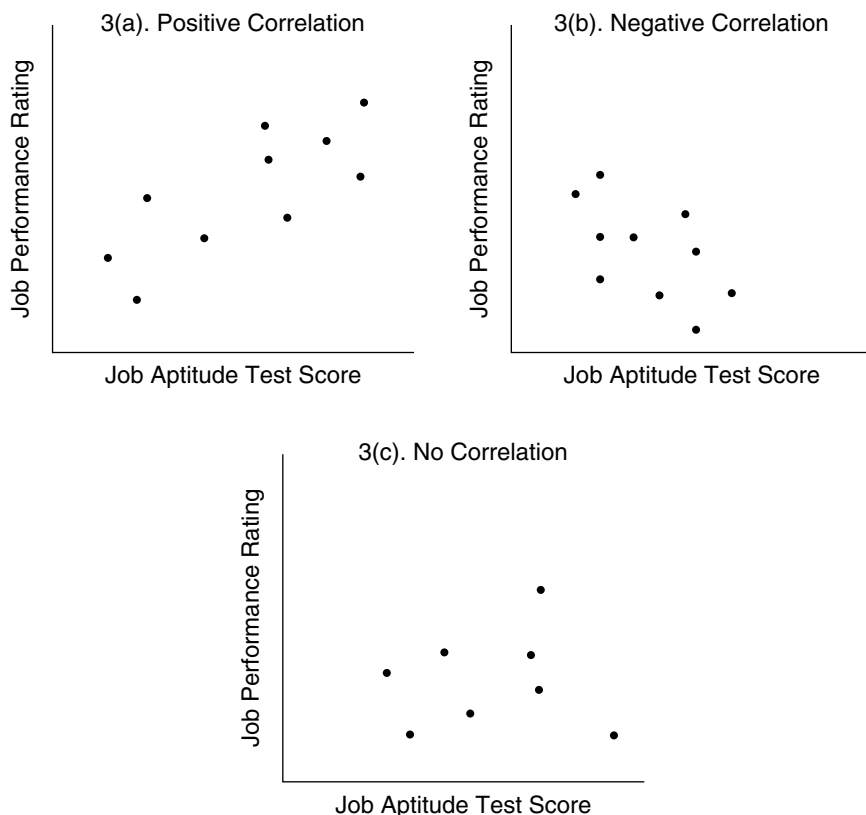
Multiple regression analysis goes beyond simple correlation by deriving the “best fitting” linear function of all the covariates in predicting the dependent variable. For example, if average job performance ratings depend on aptitude test scores, age, and education, multiple regression analysis can use information about the joint behavior of job performance and the three explanatory variables in the observed sample to derive the best fitting linear function of aptitude test scores, age, and education in predicting job performance.

Many aspects of regression analysis can be illustrated in the case where there is only a single explanatory variable. A linear relationship between an explanatory variable X and a dependent variable Y is just the equation of a line:

$$Y = a + bX \quad (1)$$

In this equation, a is the intercept of the line (i.e., the height of the line when it intersects the y -axis with $X = 0$), and b is the slope (the change in the dependent variable associated with a 1-unit change in the explanatory variable). It is important to note the “1-unit change” interpretation, because sometimes the units of measurement change (e.g., from ounces to pounds), and in that case the

Figure 3. Correlation between the job aptitude variable and the job performance variable: (a) positive correlation, (b) negative correlation, (c) weak relationship with no correlation.



meaning of a “1-unit change” also changes. For example, if b is the slope when X is measured in pounds, then the slope when X is measured in ounces will be $b/16$, since each ounce is only $1/16$ of a pound.

Unless Y is perfectly linearly related to X , equation (1) does not fully describe the determination of Y . Instead, there is a third term in the equation representing the part of Y that is not attributable to the linear effect of X :

$$Y = a + bX + \varepsilon \quad (2)$$

The term ε , usually called the residual term or the error term, incorporates all the determinants of Y apart from the intercept (also called a *constant term*) and

the linear effect of X . In many cases, an analyst will assume that the average value of the residual term is 0 for each value of X . In that case, the fact that there is a high or low value of X has no bearing on the likely sign (positive or negative) or magnitude of the other unobserved determinants of Y . As discussed above, however, in the interpretation of regression models, one of the critical issues is the plausibility of this assumption.

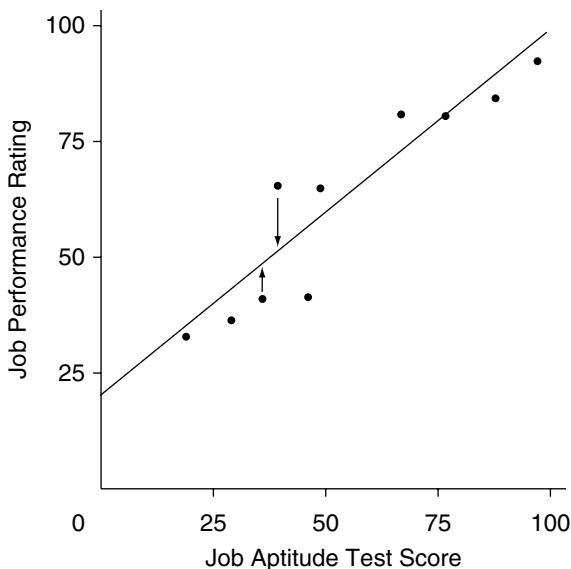
A linear regression model typically is estimated using the method of least squares (also known as ordinary least squares, or OLS), in which the values of the coefficients a and b are calculated so that the sum of the squared residual values (measured by the arrows shown in Figure 4) is made as small as possible. Note that by squaring the residuals, positive and negative deviations of the value of Y from its predicted value ($a + bX$) are given equal weight. Minimization of the squared residuals also means large deviations of Y from its predicted value receive substantially more weight in the determination of a and b than small deviations.

Figure 4, for example, shows the fitted regression line relating aptitude scores (the X variable) to job performance (the Y variable). When the aptitude test score is 0, the predicted (average) value of the job performance rating is the intercept, 18.4. Also, for each additional point on the test score, the job performance rating increases 0.73 units, which is given by the slope 0.73. Thus, the estimated regression line is¹³⁹

$$\hat{Y} = 18.4 + 0.73X$$

139. In discussions of estimated regression models, it is standard to denote the predicted value with a carrot (or “hat”) symbol.

Figure 4. Regression line.



Multiple Regression Model

When there are an arbitrary number of explanatory variables, the linear regression model takes the following form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon \quad (3)$$

where Y represents the dependent variable, such as the salary of an employee, and $X_1 \dots X_k$ represent the explanatory variables (e.g., education, years of work experience, location in a major metropolitan area, etc.). As in equation (2) above, the error term ε represents the collective influence of all omitted factors influencing Y . Sometimes an analyst will assume that these factors are purely random and can be represented as independent observations drawn from a normal distribution with mean 0 and some standard deviation. Other times, an analyst will leave the precise distribution of ε unspecified.

In a multiple regression, each of the explanatory variables has its own coefficient (e.g., β_1 for the first variable, β_2 for the second).¹⁴⁰ In addition, there is a constant or intercept term, β_0 , which has the interpretation of the expected value (or mean) of Y when all of the explanatory variables have a value of 0 (similar to the interpretation of the constant term a in equation (2) above). The full set of coefficients (also referred to as the parameters of the regression model) is usually estimated by ordinary least squares, which again selects the values to minimize the sum of the squared residuals.

Each coefficient β_k measures how the dependent variable Y responds, on average, to a change in the corresponding covariate X_k , holding constant the effects of the other covariates. This can be seen mathematically from equation (3): When all other covariates are held constant, and the value of the residual term is also held constant, a 1-unit increase in the k^{th} covariate will change the value of the dependent variable by an amount β_k .

An important feature of the use of ordinary least squares to estimate the coefficients in equation (3) is that one can give a precise interpretation of how the estimated coefficients are obtained in such a way as to reflect the “holding constant” of the other covariates. Consider the following three-step procedure. First, calculate the residuals from a regression of Y on all covariates other than X_k . These residuals reflect the part of Y that cannot be explained by the other control variables. Second, calculate the residuals of a regression of X_k on all the other covariates. These residuals reflect the part of X_k that cannot be explained by the other control variables. Third and finally, regress the first residual variable on the second residual variable. The resulting coefficient will be identical to β_k . Thus the coefficient in a multiple regression represents the slope of the line “ Y , adjusted for all covariates other than X_k versus X_k adjusted for all the other covariates.”¹⁴¹

Most statisticians and applied economists use the least-squares regression technique because of its simplicity and its desirable statistical properties. As a result, it is also used frequently in legal proceedings.

140. The variables themselves can appear in many different forms. For example, Y might represent the logarithm of an employee’s salary, and X_1 might represent the logarithm of the employee’s years of experience. The logarithmic representation is appropriate when Y increases exponentially as X increases—for each unit increase in X , the corresponding increase in Y becomes larger and larger. For example, if an expert were to graph the growth of the U.S. population (Y) over time (t), the following equation might be appropriate: $\log(Y) = \beta_0 + \beta_1 \log(t)$.

141. In econometrics, this is known as the Frisch-Waugh-Lovell theorem. Note that if X_k can be perfectly predicted by the other explanatory variables, then there will be no residuals to take to the third step—this is the extreme case of perfect multicollinearity.

Specifying the Regression Model

Suppose an expert wants to analyze the salaries of women and men at a large publishing house to discover whether a difference in salaries between employees with similar years of work experience provides evidence of discrimination.¹⁴² To begin with the simplest case, the salary in dollars per year, Y , represents the dependent variable to be explained, and X_1 represents the number of years of experience of the employee explanatory variable. The regression model would be written

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon \quad (4)$$

In equation (4), β_0 and β_1 are the parameters to be estimated from the data, and ε is the error term. The parameter β_0 is the average salary of all employees with no experience. The parameter β_1 measures the average effect of an additional year of experience on the average salary of employees.

Fitted Regression Line

Once the parameters in a regression equation have been estimated, the fitted values for the dependent variable can be calculated. If the estimated regression parameters, or regression coefficients, for the model in equation (3) are denoted as $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k$, the fitted values for Y , denoted $\hat{Y}$, are

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \dots + \hat{\beta}_k X_k \quad (5)$$

Figure 5 illustrates this for the example involving a single explanatory variable. The data are shown as a scatter of points; salary is on the vertical axis, and years of experience is on the horizontal axis. The estimated regression line is drawn through the data points. It is given by

$$\hat{Y} = \$15,000 + \$2,000X_1 \quad (6)$$

Thus, the fitted value for the salary of individual i , who has X_{1i} years of experience, will be given by

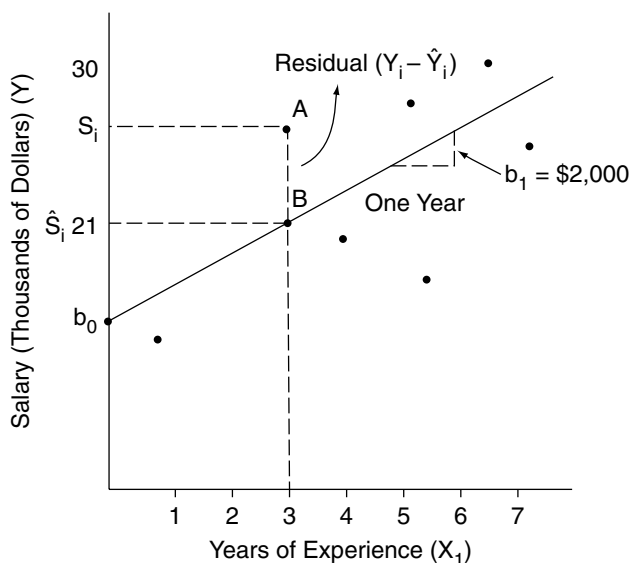
$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} \text{ (Point B)} \quad (7)$$

142. The regression results used in this example are based on data for 1,715 men and women; these results were used by the defense in a sex-discrimination case against the *New York Times* that was settled in 1978. Professor Orley Ashenfelter, Department of Economics, Princeton University, provided the data.

The intercept of the regression model represents the average value of the dependent variable when the explanatory variable or variables are equal to 0; the estimated intercept b_0 is shown on the vertical axis in Figure 5. Similarly, the slope of the line measures the (average) change in the dependent variable associated with a unit increase in an explanatory variable; the estimated slope b_1 also is shown. In equation (6), the intercept \$15,000 indicates that employees with no experience earn \$15,000 per year. The slope parameter implies that each year of experience adds \$2,000 to an “average” employee’s salary.

Now, suppose that the salary variable is related simply to the sex of the employee. The relevant indicator variable, often called a dummy variable, is X_2 , which is equal to 1 if the employee is male, and 0 if the employee is female. Suppose the regression of salary Y on X_2 yields the following result: $Y = \$30,449 + \$10,979X_2$. The coefficient \$10,979 measures the difference between the average salary of men and the average salary of women.¹⁴³

Figure 5. Goodness of fit.



143. To understand why, note that when $X_2 = 0$, the average salary for women is $\$30,449 + \$10,979 \times 0 = \$30,449$. Correspondingly, when $X_2 = 1$, the average salary for men is $\$30,449 + \$10,979 \times 1 = \$41,428$. The difference— $\$41,428 - \$30,449$ —is \$10,979.

Regression residuals

Recall that for each data point, the regression residual is the difference between the actual value of the dependent variable and part of that value that is attributable to the explanatory variables. After a regression model has been estimated, the estimated regression residual ($\hat{\varepsilon}$) is the difference between the actual values and the fitted or predicted values from the estimated model. Suppose, for example, that we are studying an individual with three years of experience and a salary of \$27,000. According to the estimated regression line in Figure 5, the predicted salary of an individual with three years of experience is \$21,000. (This can also be interpreted as the expected salary that an individual with three years of experience would receive, since all the unexplained terms in the residual should on average take a value of 0.) Because the individual's salary is \$6,000 higher than the predicted salary from the model, the estimated residual (the individual's salary minus the predicted salary, based on the estimated coefficients) is \$6,000. In general, the estimated residual $\hat{\varepsilon}$ associated with a data point, such as Point A in Figure 5, is given by $\hat{\varepsilon}_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i}$. Each data point in the figure has an estimated residual, which is the error made by the least-squares regression method for that individual.

Interaction terms

Models with interaction terms allow for the possibility that the effect of an explanatory variable on the dependent variable may vary in magnitude as the level of other explanatory variables changes. As a starting point, assume that the regression model takes the following form:

$$Y = \beta_0 + \beta_1 \text{MALE} + \beta_2 \text{EXP} + \varepsilon \quad (8)$$

where Y is annual salary, MALE is equal to 1 for men and 0 for women, EXP represents years of job experience, and ε is the error term. The coefficient β_1 measures the difference in average salary (across all experience levels) between men and women. In essence the inclusion of a dummy variable such as MALE allows for a parallel vertical shift in the regression line. The shift is parallel because the slope coefficient β_2 , which measures the effect of experience on salary, is the same for both males and females.

Now suppose that the regression model is expanded to take on the following form:

$$Y = \beta_0 + \beta_1 \text{MALE} + \beta_2 \text{EXP} + \beta_3 \text{EXP} \times \text{MALE} + \varepsilon \quad (9)$$

In this model, the coefficient β_1 measures the difference in average salary (across all experience levels) between men and women for employees with no experience. The coefficient β_2 measures the effect of experience on salary for women (when MALE = 0), and the interaction coefficient β_3 measures the difference in the effect of experience on salary between men and women. It follows, for example, that the effect of one year of experience on salary for women is β_2 , whereas the comparable effect for men is $\beta_2 + \beta_3$.¹⁴⁴

Interpreting regression results

To explain how regression results are interpreted, we can expand the earlier example associated with Figure 5 to consider the possibility of an additional explanatory variable—the square of the number of years of experience, X_3 . The X_3 variable is designed to capture the fact that for most individuals, salaries increase with experience, but at a faster rate for people who first enter the labor market (with low levels of experience), and a lower rate for people with more experience. Thus, we might expect the estimated coefficient of X_1 to be positive and the estimated coefficient of X_3 to be negative.¹⁴⁵

The estimated regression line using the third additional explanatory variable, as well as the first explanatory variable for years of experience (X_1) and the dummy variable for male gender (X_2), is

$$\hat{Y} = \$14,085 + \$2,323X_1 + \$1,675X_2 - \$36X_3$$

The importance of including relevant explanatory variables in a regression model is illustrated by the change in the regression results after the X_3 and X_1 variables are added. The coefficient on the variable X_2 measures the difference in the salaries of men and women while controlling for the linear and quadratic effects of experience. The differential of \$1,675 is substantially lower than the previously measured differential of \$10,979. Clearly, failure to control for job experience in this example leads to an overstatement of the difference in salaries between men and women.

Now consider the interpretation of the explanatory variables for experience, X_1 and X_3 . The positive sign on the X_1 coefficient shows that salary increases with experience. The negative sign on the X_3 coefficient indicates that the rate of

144. Estimating a regression in which there are interaction terms for all explanatory variables, as in equation (9), is essentially the same as estimating two separate regressions, one for men and one for women.

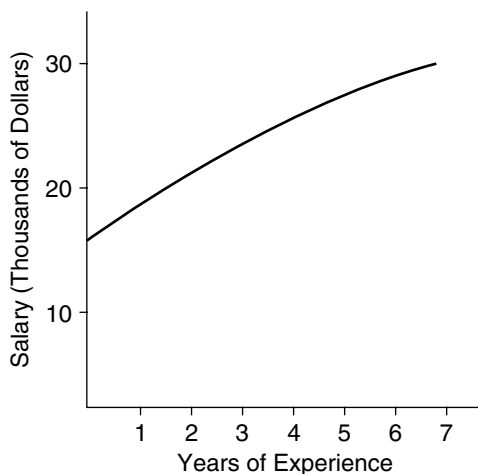
145. A model that includes a variable and its square is often referred to as a *quadratic* specification. Thus, the equation on this page would be referred to as a model that includes controls for “a quadratic in experience.”

salary increase diminishes with experience (as expected). To determine the combined effect of the variables X_1 and X_3 , some simple calculations can be made.¹⁴⁶ For example, consider how the average salary of women ($X_2=0$) changes with the level of experience. As experience increases from zero to one year, the average salary increases by \$2,287 (\$2,323 – \$36), from \$14,085 to \$16,372. However, women with two years of experience earn only \$2,215 more than women with one year of experience, and women with three years of experience earn only \$2,143 more than women with two years.¹⁴⁷ Figure 6 illustrates the results: The regression line shown is for women’s salaries; the corresponding line for men’s salaries would be parallel and \$1,675 higher.

The problem of omitted variables

A major problem confronting the interpretation of regression models is the possibility that some of the unobserved or omitted factors included in the error term are in fact correlated with the included variables. For example, consider an analysis of the effect of having attended a prestigious law school on the earnings of attorneys 10 years after completion of their law degrees. The available dataset

Figure 6. Regression slope for women’s salaries.



146. More formally, consider a model that relates a dependent variable to some variable Z with a coefficient of β_1 and to its square (Z^2) with a coefficient of β_2 (i.e., $Y = \beta_1 Z + \beta_2 Z^2 + \dots$). The marginal effect of a small increase in Z on the predicted value of the dependent variable is $\beta_1 + 2\beta_2 Z$, which is the derivative of the quadratic expression.

147. These numbers can be calculated by substituting different values of X_1 and X_3 in the previous equation.

includes a sample of students who attended different law schools, a measure (Y) of their annual salary in the tenth year post-graduation, an indicator (X_1) for having attended a “top 20” law school; an indicator for female gender (X_2), and an indicator (X_3) for whether their undergraduate degree was in a STEM (Science, Technology, Engineering, and Mathematics) field or not. The expert intends to interpret the coefficient on attending a top 20 school as the causal effect of the superior training and network opportunities afforded by such a school, as part of litigation over the admission rules used by certain schools.

The proposed regression model takes the following form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon \quad (10)$$

Included in the error term ε are all the other unmeasured factors affecting salary, apart from the three covariates. Some of these factors, such as interest in pursuing a public interest law career, may negatively affect salary, while others, like ambition or ability, may positively affect salary. Moreover, it is likely that the average value of ε is different for students who attended a top 20 program versus those who did not. How does this affect the estimates of the coefficient β_1 ?

It turns out there is a very simple answer, at least conceptually. Consider the *hypothetical* regression model relating ε to the three covariates:

$$\varepsilon = \lambda_0 + \lambda_1 X_1 + \lambda_2 X_2 + \lambda_3 X_3 + \xi \quad (11)$$

Here, the coefficients λ_1 , λ_2 , and λ_3 represent the effects of attending an elite school, being female, or having a STEM undergraduate degree on the combination of omitted factors that are included in ε . For example, if $\lambda_1 > 0$, then attending an elite school is associated, on average, with a more positive combination of omitted factors. Note that if the error term in equation (10) is truly random, and independent of the included covariates, then all the λ coefficients in (11) will be zero. Alternatively, if X_1 were randomly determined by an admissions lottery, then λ_1 would be zero, since there is no information in a randomized variable that can be used to predict other variables.

When the regression model (10) is estimated by OLS, the coefficients of the included covariates will incorporate both the true causal effects of the covariates (i.e., the β s), and the effect of the covariates in predicting ε (i.e., the λ s):

$$Y = (\beta_0 + \lambda_0) + (\beta_1 + \lambda_1)X_1 + (\beta_2 + \lambda_2)X_2 + (\beta_3 + \lambda_3)X_3 + \xi \quad (12)$$

Notice that the error in this model is the error from the hypothetical regression model (11): this is the part of ε that cannot be predicted by the included covariates. Equation (12) says that in estimating a multiple regression of salaries on elite school status, gender, and STEM major, the coefficient associated with attending an elite school is actually $\beta_1 + \lambda_1$, reflecting the fact that attending an

elite law school has a direct effect on salary, β_1 , holding constant the other controls and all the unobservable factors in ε , plus an indirect effect arising because on average people who attend an elite law school have different values of ε . This second term, λ_1 , is known as omitted-variables bias.

To summarize, if one estimates a multiple regression model relating an outcome to a set of observed covariates, the estimated coefficients will be estimates of the combined effect of the causal coefficients in the intended model *and* the coefficients that arise in trying to predict whatever is left in the error term in this model. This happens because the actual regression coefficients are algorithmically selected to form the best possible estimate of Y , given the observed covariates. In choosing the coefficients, the OLS algorithm *does not distinguish* between the intended causal effects in the model and the potential role of the covariates in predicting the residual component in the model.

Although equations (10)–(12) provide a useful framework for thinking about omitted-variables biases, it is important to keep in mind that the hypothetical regression (11) cannot be estimated, because we cannot actually see the value of ε . Nevertheless, analysts often present detailed discussions of the likely sign (and less commonly, the magnitude) of the coefficients in the hypothetical regression. If it can be argued, for example, that λ_1 is almost surely positive, then one can infer that the estimated effect of attending an elite law school from a model based on (10) is almost surely upward biased as an estimate of β_1 . In addition, sometimes information from another sample can be used to help assess the signs and magnitudes of omitted-variables biases. For example, suppose information from a previous study shows that higher LSAT scores are associated with higher salaries and that information is available on the average LSAT scores of students attending different law schools. From these sources it might be possible to make an educated guess about a plausible magnitude for λ_1 .

Causal Modeling

Pre/Post Comparisons

Concerns over omitted-variables bias in regression models have led to the development of alternative regression-based approaches for estimating the causal effect of a variable of interest. These approaches typically focus on situations where the value of the explanatory variable of interest changes discretely *for a known reason* (e.g., a political decision).¹⁴⁸ For example, suppose one is interested in understanding the effect of more intensive policing on crime. One might then focus on cases where a new policing program (e.g., New York City’s “stop and frisk” program) is introduced. If data on crime rates are available from both *before* and *after*

148. Sometimes these changes are referred to as *quasi-experiments* or *natural experiments*.

the introduction, a pre/post estimate of the effect of the policy can be constructed from the change in the average number of crimes per day following the implementation of the new policy.

In a pre/post research design (and in related difference-in-differences designs), it is necessary to distinguish between the two separate dimensions over which the data are observed. One dimension is across different units, indexed by the subscript i (e.g., in a study of policing and crime, the units of observation may be police precincts); the other is over time (e.g., different months or years), indexed by the subscript t . Using this notation, let Y_{it} represent the outcome of interest observed for unit i in period t . Assume that a new program or intervention is initiated at time $t=1$. In earlier periods ($t=0$, $t=-1$, $t=-2$, . . .), the program was absent; in all periods on or after $t=1$, the program is in place. A very simple model for Y_{it} is one that includes a constant and an indicator for post-intervention periods, $Post_t = 0$ if $t \leq 0$, and $Post_t = 1$ if $t \geq 1$:

$$Y_{it} = \beta_0 + \beta_1 Post_t + \varepsilon_{it}. \quad (13)$$

This simple model implies that in all periods up to period 0,

$$Y_{it} = \beta_0 + \varepsilon_{it} \quad (14)$$

whereas in all later periods $t=1, 2, \dots$,

$$Y_{it} = \beta_0 + \beta_1 + \varepsilon_{it}. \quad (15)$$

Setting $t=1$ in equation (15) and setting $t=0$ in equation (14) and subtracting, the change in the outcome for unit i from period 0 to period 1 is

$$\Delta Y_i = Y_{i1} - Y_{i0} = \beta_1 + \varepsilon_{i1} - \varepsilon_{i0}. \quad (16)$$

Treating $\varepsilon_{i1} - \varepsilon_{i0}$ as a residual, equation (16) is the simplest possible regression model, with a constant term and no other covariates. The OLS estimate of β_1 in such a model is exactly the mean change in the outcome across all the units in the sample, that is,

$$\hat{\beta}_1 = \text{Mean}(\Delta Y_i) \quad (17)$$

where $\text{Mean}(Y)$ denotes the mean value of the variable Y in the sample.

To interpret equation (17), note that $\text{Mean}(\Delta Y_i) = \text{Mean}(Y_{i1}) - \text{Mean}(Y_{i0})$. In a pre/post design, the mean outcome in the post period is compared to the mean outcome in the pre period. If the mean value of the outcome does not change after the intervention, the estimated effect of the intervention is 0. If the mean outcome rises (falls), then $\hat{\beta}_1$ will be positive (negative). Conceptually, a pre/post

design is using the mean of the outcome in the pre-intervention period as an estimate of what the mean *would have been* in the absence of the intervention. This is known as the counterfactual mean in applied economics, or the but-for mean in legal discussions. The difference between the actual mean in the post-intervention period and the counterfactual mean is the estimated effect of the intervention.

If data were only available from the period after the start of the intervention, an alternative approach would be to try to find another set of units that have not adopted the intervention and use the mean of the outcomes for this “comparison group” as a counterfactual. In the policing example, suppose that some precincts adopt the new program, and some do not. Then one might consider using the average crime rates for precincts that did not adopt the policy as the counterfactual for those that did.¹⁴⁹ A concern with this alternative approach is that there may be differences between units that adopted the new program and units that did not, raising the possibility of omitted-variables bias. In a pre/post design, the counterfactual is based on the average outcomes *of the same units* before they implemented the intervention. In some cases, this counterfactual will be more compelling.

The potential advantages of a pre/post design are illustrated by decomposing ε_{it} , the error term in equation (13), into two parts: a part α_i that is constant over time, representing the mean value of all the ε_{it} 's for unit i , and the deviation of ε_{it} from its mean value in period t ,

$$\varepsilon_{it} = \alpha_i + \nu_{it} \quad (18)$$

For example, in a study of crime and policing where i indexes different precincts and t has two values, $t=0$ and $t=1$, the error ε_{it} from the model in equation (11) has the interpretation of the deviation in the crime rate in a precinct i in period t from the average crime rate in the city in the same period. One might expect certain precincts to have higher average crime rates in both periods, and others to have lower crime rates. Such differences in average rates would be reflected in the α_i component of equation (18), whereas unexplained fluctuations in crime from period to period in the same precinct would be reflected in the ν_{it} component. From equation (18) it follows that the difference in ε_{it} between period 0 and period 1, which is the error term in equation (16), is

$$\varepsilon_{i1} - \varepsilon_{i0} = \nu_{i1} - \nu_{i0} \quad (19)$$

Differencing eliminates the time-invariant component of ε_{it} , leaving a new error term that is less likely to be affected by omitted-variable biases, since all the

149. This would replace equation (13) with an alternative model: $Y_i = \beta_0 + \beta_1 \text{Adopt}_i + \varepsilon_i$, where Adopt_i is an indicator variable equal to 1 for precincts that adopted the policy and 0 for those that did not. In such a model, the OLS estimate of the coefficient β_1 is $\hat{\beta}_1 = \text{Mean}(Y | \text{Adopt}_i = 1) - \text{Mean}(Y | \text{Adopt}_i = 0)$, where $\text{Mean}(Y|C)$ denotes the mean in the sample among units for which condition C is true.

permanent factors that differ across units are removed through the differencing process.

Nevertheless, a simple pre/post design has an important limitation: *any change* that affects the average value of the outcome between period 0 and period 1 is attributed to the intervention. In many situations, the mean value of an outcome changes over time for reasons that have nothing to do with the intervention or change in policy under study. One way to check whether such factors are present is to construct average differences in the outcome for earlier periods and verify that these are all very close to zero. For example, equation (13) implies that the *lagged difference* in outcomes $\Delta Y_i^{(-1)}$ is determined by

$$\Delta Y_i^{(-1)} = Y_{i0} - Y_{i(-1)} = \varepsilon_{i0} - \varepsilon_{i(-1)}. \quad (20)$$

The same arguments that support the contention that the error term $\varepsilon_{i1} - \varepsilon_{i0}$ in equation (16) has a mean value of 0, so that $Mean(\Delta Y_i)$ provides a good estimate of β_1 , imply that the error term in equation (20) also has a mean value of 0. Thus, $Mean(\Delta Y_i^{(-1)})$ should be very close to 0. Indeed, the mean values of the changes in the outcome for all periods where there was no change in policy should be close to 0. In the absence of evidence that this is true, a simple pre/post design is not very credible.

Difference in Differences

A natural way to address the presence of unmeasured factors that are changing over time and confounding the interpretation of a simple pre/post design is to compare the change in the outcome for units where the new program was adopted to the change in the outcome over the same period for units that did not adopt the program. This is the difference-in-differences approach.

To spell this out, call the treatment group the set of units where the program is adopted between period 0 and period 1; call the other set of units the comparison group, and let T_i be an indicator variable that is equal to 1 for units in the treatment group. Suppose that data are only observed in periods 0 and 1. Then, the difference-in-differences model can be written as a multiple linear regression model of the form

$$Y_{it} = \beta_0 + \beta_1 T_i \times Post_t + \beta_2 T_i + \beta_3 Post_t + \varepsilon_{it} \quad (21)$$

The key coefficient in this model is β_1 , the coefficient on the interaction between the indicator for the treatment group and the indicator for period 1 (the post period).

To understand this equation, notice that for units with $T_i = 0$, it implies

$$Y_{it} = \beta_0 + \beta_3 Post_t + \varepsilon_{it} \quad (22)$$

Thus, the change in outcomes for units in the comparison group is

$$\Delta Y_i = Y_{i1} - Y_{i0} = \beta_3 + \varepsilon_{i1} - \varepsilon_{i0} \quad (23)$$

The inclusion of the term $\beta_3 Post_t$ in equation (22) captures the possibility that even in the comparison group, unobserved factors can change between period 0 and period 1. Assuming the mean value of $\varepsilon_{i1} - \varepsilon_{i0}$ is 0 for comparison units, the expected average change in the outcome in the comparison group will be β_3 .¹⁵⁰ For units in the treatment group, equation (21) implies

$$Y_{it} = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)Post_t + \varepsilon_{it} \quad (24)$$

Thus, the change in outcomes for units in the treatment group is

$$\Delta Y_i = (\beta_1 + \beta_3) + \varepsilon_{i1} - \varepsilon_{i0} \quad (25)$$

Assuming that the mean value of $\varepsilon_{i1} - \varepsilon_{i0}$ is 0 for units in the treatment group, the expected average change in the outcome in the treatment group will be $\beta_1 + \beta_3$, which is a combination of the change experienced by the comparison group, β_3 , and the effect of the program, β_1 . In fact, the OLS estimate of β_1 in the difference-in-differences model (21) is exactly

$$\hat{\beta}_1 = Mean(\Delta Y_i | T_i = 1) - Mean(\Delta Y_i | T_i = 0) \quad (26)$$

where $Mean(\Delta Y_i | T_i = 1)$ denotes the mean in the subsample with $T_i = 1$ (i.e., among the treatment group), and $Mean(\Delta Y_i | T_i = 0)$ denotes the mean in the subsample with $T_i = 0$ (i.e., among the comparison group).

The critical assumption in a difference-in-differences design is that in the absence of the program, the mean change in outcomes for the treatment group and comparison groups would have been the same. Under that assumption, the mean change for the comparison group forms a counterfactual for the change that would have occurred in the treatment group in the absence of the program. As noted previously, this assumption is sometimes referred to as a parallel trends assumption. With only two periods of data, it cannot be directly evaluated. If there are additional periods of data, however, parallel trends can be assessed by forming differences in differences for other periods that do not overlap with the start of the program. These should all be close to zero. Visually, a graph of the mean outcomes for the treatment group over a sequence of periods (for example, $t = -2, -1, 0, 1, 2$) should show that the the gap between the treatment means and

150. To be precise, the “expected average change” means the mathematical expectation of the sample average change in the outcome.

comparison means is stable prior to the program date, then shifts discretely between period 0 and 1, reflecting the treatment effect, then stabilizes again.

Another way of thinking about the validity of a difference-in-differences specification is to ask whether the choice of the pre-treatment period matters. For example, consider an evaluation of a state law that liberalizes state divorce law. A difference-in-differences approach to measuring the effect of the law would be to compare the change in divorce rates from some year prior to the law to the year after the law changed in the “treatment” state (the state that passed the new law) relative to the change in divorce rates in a “comparison” state. If the assumptions of a difference-in-differences specification are correct, then the relative change in divorce rates in the treated state relative to the comparison state should not depend on the particular year that is selected as the benchmark prior to the law change. Under those assumptions, the difference in divorce rates in the treated and comparison states can be expected to be similar in all the years prior to the new law. As a result, choosing one particular year as a benchmark versus another will not matter. Graphically, this requires that the year-to-year changes in divorce rates in the treated state and the comparison state are the same, and a plot of the divorce rates prior to the law change would appear as two parallel lines. This is the origin of the term *parallel trends*.

A common scenario where the parallel trends assumption may fail is one in which the timing of the program is correlated with the trend in recent outcomes for the treatment group. For example, Ashenfelter (1978) noted that people who entered government retraining programs had *temporarily depressed* earnings in the period before entry, driven by recent job losses or other problems—the Ashenfelter dip.¹⁵¹ Building on equation (18), this suggests that program participants may be particularly likely to have large negative values of ν_{i0} (the period-specific part of their earnings residual in the period just prior to the program). In that case, the mean value of the error term $\varepsilon_{i1} - \varepsilon_{i0} = \nu_{i1} - \nu_{i0}$ in equation (25) will be positive, implying that

$$Mean(\Delta Y_i | T_i = 1) = \beta_1 + \beta_3 + Mean(\varepsilon_{i1} - \varepsilon_{i0} | T_i = 1) > \beta_1 + \beta_3 \quad (27)$$

causing a difference-in-differences estimator to overstate the effect of the program. The possibility of such pre-trends can be evaluated empirically by examining the mean outcomes of the treatment and comparison groups in the periods before the start of the program and making sure that the parallel trends assumption is valid.

151. Ashenfelter, *supra* note 94.

Instrumental Variables

An alternative method that is widely used to address omitted variables problems in regression models is instrumental variables (IV). Consider the case where an expert has posited a model relating an outcome Y to an observed explanatory variable:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad (28)$$

For clarity, subscripts indicate the different units (indexed by i) in the sample. The expert wants to give a causal interpretation of the coefficient β_1 . Specifically, in the expert's model, β_1 represents the effect of a unit change in X on the average or expected value of the outcome, holding constant all other unmeasured/unobserved determinants of Y . Such an equation is known as the structural model in IV and related settings. The problem is that, if this structural equation is estimated by OLS, the estimate $\hat{\beta}_1$ will incorporate both the desired causal effect and any potential omitted variable bias that arises because X_i is potentially correlated with some of the factors included in the error ε_i . This omitted-variable bias is sometimes called a simultaneity bias or endogeneity bias—particularly in cases where there is feedback from Y_i to X_i , as discussed in the section titled “Instrumental Variables Design” above. Whatever the source of the potential correlation between X_i and ε_i , however, it can be analyzed in the omitted variables framework of equations (10)–(12).

The objective of IV estimation is to isolate the part of the variation in X_i attributable to variation in some other variable, Z_i (the instrumental variable or the instrument), and use only that variation in estimating the effect of X_i on Y_i . An appropriate instrumental variable has to satisfy three assumptions: (1) it has an effect on X_i , so there is a component of X_i that can be attributed to Z_i ; (2) it is uncorrelated with ε_i , the unobserved determinants of Y_i ; and (3) it has no direct effect on the outcome Y_i . An example of an instrumental variable that satisfies these requirements is a randomly assigned lottery outcome. For example, in a study of the effect of attending a charter school on test scores, the analysis showed that if access to over subscribed charter schools is determined by lottery, then an appropriate instrumental variable is an indicator for whether a student was assigned to the charter school by the admission lottery.¹⁵²

The IV method involves estimating two regression models. The first-stage model is a linear regression relating the variable of interest (X_i) to the instrumental variable:

$$X_i = \pi_0 + \pi_1 Z_i + u_i \quad (29)$$

152. See generally Chabrier et al., *supra* note 104.

From this model it is possible to form the predicted values of X_i based only on Z_i :

$$\hat{X}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i \quad (30)$$

The second-stage model is a linear regression of the outcome on the predicted values of X_i :

$$Y_i = \beta_0^{iv} + \beta_1^{iv} \hat{X}_i + \eta_i \quad (31)$$

The OLS estimate of the coefficient of $\hat{X}_i$ in the second-stage model, $\hat{\beta}_1^{iv}$, is the IV estimate of β in the structural model (28). Note that since $\hat{X}_i$ only depends on Z_i , the second-stage equation measures the effect of the part of $\hat{X}_i$ that is determined by Z_i on the outcome variable.

The IV estimate can be derived in another way. Substituting the first-stage equation (29) into the structural equation (28) yields an equation relating Y to Z :

$$Y_i = \underbrace{(\beta_0 + \beta_1 \pi_0)}_{\delta_0} + \underbrace{\beta_1 \pi_1}_{\delta_1} Z_i + \underbrace{(\varepsilon_i + \beta_1 u_i)}_{\xi_i} = \delta_0 + \delta_1 Z_i + \xi_i \quad (32)$$

This equation, known as the reduced-form model, can be estimated by OLS without concern about omitted variables bias because, under the assumption that Z_i is a valid instrumental variable, it is uncorrelated with ξ_i .¹⁵³ But equation (32) shows that the coefficient of the instrumental variable in the reduced-form model is $\delta_1 = \beta_1 \pi_1$. This occurs because under assumption #3 for a valid instrumental variable, the *only way* that Z_i affects Y_i is indirectly, through its effect on X_i . Since Z_i affects X_i with a coefficient of π_1 in the first-stage model, and X_i affects Y_i with a coefficient of β_1 in the structural model, the reduced-form effect of Z_i on Y_i is the product of these two coefficients. This reasoning suggests that one could obtain an estimate of β_1 by dividing the estimated reduced-form coefficient, $\hat{\delta}_1$, by the estimated first-stage coefficient $\hat{\pi}_1$. And in fact this ratio is exactly the IV estimate:

$$\hat{\beta}_1^{iv} = \frac{\hat{\delta}_1}{\hat{\pi}_1} \quad (33)$$

The expression for the IV estimate in equation (33) is particularly insightful in the case where X_i is a dummy variable indicating participation in a program or

153. Under assumption #2 for a valid instrumental variable, Z_i is uncorrelated with ε_i . Moreover, the error in the first-stage equation u_i is necessarily uncorrelated with the explanatory variable in that model, by a fundamental property of least squares. Thus Z_i is uncorrelated with the composite error in the reduced-form model, $\xi_i = \varepsilon_i + \beta_1 u_i$.

intervention (for example, enrollment in a charter school), and Z_i is also a dummy variable, indicating a condition that partly determines program participation (for example, whether student i was assigned to the charter school by lottery). In this case, the first-stage coefficient $\hat{\pi}_1$ is the difference in program participation rates between units with Z_i equal to 0 or 1:

$$\hat{\pi}_1 = \text{Mean}(X_i | Z_i = 1) - \text{Mean}(X_i | Z_i = 0), \quad (34)$$

and the reduced-form coefficient $\hat{\delta}_1$ is the difference in the mean of the outcome variable (e.g., an end-of-year test score) for the same two groups:

$$\hat{\delta}_1 = \text{Mean}(Y_i | Z_i = 1) - \text{Mean}(Y_i | Z_i = 0). \quad (35)$$

Thus, equation (33) implies that the IV estimator for the effect of program participation on the outcome is

$$\hat{\beta}_1^{iv} = \frac{\text{Mean}(Y_i | Z_i = 1) - \text{Mean}(Y_i | Z_i = 0)}{\text{Mean}(X_i | Z_i = 1) - \text{Mean}(X_i | Z_i = 0)} \quad (36)$$

The IV estimator takes the difference in mean outcomes between the group with $Z_i = 1$ and the group with $Z_i = 0$, and divides this by the difference in the fraction who participated in the program when Z_i was 1 or 0. Intuitively, the IV estimator is simply creating a *per unit* estimate of the treatment effect of the intervention, based on the differences between groups with different values of Z_i .

Determining the Precision of the Regression Results

Least squares regression algorithms provide not only parameter estimates that indicate the direction and magnitude of the effect of a change in the explanatory variable on the dependent variable, but also an estimate of the reliability of the parameter estimates and a measure of the overall goodness of fit of the regression model. Each of these factors is considered in turn.

Standard Errors of the Coefficients and t-Statistics

Estimates of the true but unknown parameters of a regression model are numbers that depend on the sample of observations under study. If a different sample

were used, a different estimate would be calculated.¹⁵⁴ If the expert continued to collect more and more samples and generated additional estimates, as might happen when new data became available over time, the estimates of each parameter would follow a probability distribution. This probability distribution can be summarized by a mean and a measure of dispersion around the mean, a standard deviation, which usually is referred to as the standard error of the coefficient, or the standard error (SE).¹⁵⁵

Suppose, for example, that an expert is interested in estimating the average price paid for a gallon of unleaded gasoline by consumers in a particular geographic area of the United States at a particular point in time. The mean price for a sample of 10 gas stations might be \$1.25, while the mean for another sample might be \$1.29, and the mean for a third, \$1.21. On this basis, the expert also could calculate the overall mean price of gasoline to be \$1.25 and an estimate of the standard deviation to be \$0.04.

Least-squares regression generalizes this result, by calculating means whose values depend on one or more explanatory variables. The standard error of a regression coefficient tells the expert how much parameter estimates are likely to vary from sample to sample. The greater the variation in parameter estimates from sample to sample, the larger the standard error and consequently the less reliable the regression results. Small standard errors imply results that are likely to be similar from sample to sample, whereas results with large standard errors show more variability.

Under appropriate assumptions, the ordinary least-squares estimators provide “best” determinations of the true underlying parameters.¹⁵⁶ In fact, OLS has several desirable properties. First, assuming that the error term in the regression model is uncorrelated with the covariates, least squares estimators are unbiased. Intuitively, this means that if the regression were calculated repeatedly with different samples, the average of the many estimates obtained for each coefficient would be the true parameter. Second, under the same assumption, least-squares estimators are consistent; if the sample were very large, the estimates obtained would come close to the true parameters. Third, least squares is efficient, in that its estimators have the smallest variance among all (linear) unbiased estimators.

If the further assumption is made that the probability distribution of each of the error terms is known, statistical statements can be made about the precision

154. The least squares formula that generates the estimates is called the least squares estimator, and its values vary from sample to sample.

155. See David H. Kaye & Hal S. Stern, *Reference Guide on Statistics and Research Methods*, “Randomized Controlled Experiments” section, in this manual.

156. The necessary assumptions of the regression model include (a) the model is specified correctly, (b) errors associated with each observation are drawn randomly from the same probability distribution and are independent of each other, (c) errors associated with each observation are independent of the corresponding observations for each of the explanatory variables in the model, and (d) no explanatory variable is correlated perfectly with a combination of other variables.

of the coefficient estimates. For relatively large samples (often, thirty or more data points will be sufficient for regressions with a small number of explanatory variables), the probability that the estimate of a parameter lies within an interval of 1.96 standard errors around the true parameter is approximately .95, or 95%. A frequent, although not always appropriate, assumption in statistical work is that the error term follows a normal distribution, from which it follows that the estimated parameters are normally distributed. The normal distribution has the property that the area within 1.96 standard errors of the mean is equal to 95% of the total area. Note that the normality assumption is not necessary for least squares to be used, because most of the properties of least squares apply regardless of normality.

In general, for any parameter estimate $\hat{\beta}$, the expert can construct an interval around $\hat{\beta}$ such that if the sampling procedure were conducted 100 times, approximately 95 of the resulting intervals would be expected to contain the true population mean. The 95% confidence interval¹⁵⁷ is given by¹⁵⁸

$$\hat{\beta} \pm 1.96(\text{SE of } \hat{\beta}) \quad (37)$$

The expert can test the hypothesis that a parameter is equal to 0 (often stated as testing the null hypothesis) by looking at its *t*-statistic, which is defined as

$$t = \hat{\beta} / \text{SE}(\hat{\beta}) \quad (38)$$

If the *t*-statistic is less than 1.96 in magnitude, the 95% confidence interval around $\hat{\beta}$ must include 0.¹⁵⁹ Because this means that the expert cannot reject the hypothesis that β equals 0, the estimate, whatever it may be, is said to be not statistically significant. Conversely, if the *t*-statistic is greater than 1.96 in absolute value, the expert concludes that the true value of β is unlikely to be 0 (intuitively, $\hat{\beta}$ is “too far” from 0 to be consistent with the true value of β being 0). In this case, the expert rejects the hypothesis that β equals 0 and calls the estimate statistically significant. If the null hypothesis β equals 0 is true, using a 95% confidence level will cause the expert to falsely reject the null hypothesis 5% of the time. Consequently, results often are said to be significant at the 5% level.¹⁶⁰

157. Confidence intervals are used commonly in statistical analyses because the expert can never be certain that a parameter estimate is equal to the true population parameter.

158. If the number of data points in the sample is small (perhaps less than 30), the 1.96 number must be adjusted upward.

159. The *t*-statistic applies to any sample size. As the sample size increases, the underlying distribution, which is the source of the *t*-statistic (Student's *t*-distribution), approximates the normal distribution.

160. A *t*-statistic of 2.57 in magnitude or greater is associated with a 99% confidence level, or a 1% level of significance, that includes a band of 2.57 standard deviations on either side of the estimated coefficient.

As an example, consider a more complete set of regression results associated with the salary regression described earlier:

$$\begin{array}{ccccccc} \hat{Y} = \$14,085 + \$2,323X_1 + \$1,675X_2 - \$36X_3 & & & & & & \\ (1,577) & (140) & (1,435) & (3.4) & & & \\ t = 8.9 & 16.5 & 1.2 & -10.8 & & & (39) \end{array}$$

The standard error of each estimated parameter is given in parentheses directly below the estimated coefficient, and the corresponding *t*-statistics appear below the standard error values.

Consider the coefficient on the dummy variable X_2 , representing the sex of a worker. It indicates that \$1,675 is the best estimate of the mean salary difference between men and women. However, the standard error of \$1,435 is large in relation to its coefficient \$1,675. Because the standard error is relatively large, the range of possible values for measuring the true salary difference, the true parameter, is great. In fact, a 95% confidence interval is given by the range negative \$1,138 to positive \$4,488. In other words, the expert can have 95% confidence that the true value of the coefficient lies between -\$1,138 and \$4,488. Because this range includes 0, the effect of sex on salary is said to be insignificantly different from 0 at the 5% level. The *t* value of 1.2 is equal to \$1,675 divided by \$1,435. Because this *t*-statistic is less than 1.96 in magnitude (a condition equivalent to the inclusion of a 0 in the above confidence interval), the sex variable again is said to be an insignificant determinant of salary at the 5% level of significance.

Note also that experience is a highly significant determinant of salary, because both the X_1 and the X_3 variables have *t*-statistics substantially greater than 1.96 in magnitude. More experience has a significant positive effect on salary, but the size of this effect diminishes significantly with experience.

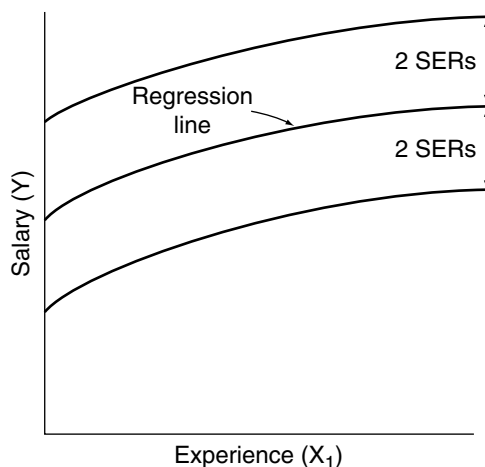
Goodness of Fit

Reported regression results usually contain not only the point estimates of the parameters and their standard errors or *t*-statistics, but also other information that tells how closely the regression line fits the data. One statistic, the standard error of the regression (SER), is an estimate of the overall size of the regression residuals.¹⁶¹ An SER of 0 would occur only when all data points lie exactly on the regression line—an extremely unlikely possibility. Other things being equal, the larger the SER, the poorer the fit of the data to the model.

161. More specifically, it is a measure of the standard deviation of the regression error ϵ . It sometimes is called the root mean square error of the regression line.

For a normally distributed error term, the expert would expect approximately 95% of the data points to lie within two SERs of the estimated regression line, as shown in Figure 7.

Figure 7. Standard error of the regression (SER).



R -squared (R^2) is a statistic that measures the percentage of variation in the dependent variable that is accounted for by all the explanatory variables.¹⁶² Thus, R^2 provides a measure of the overall goodness of fit of the multiple regression equation. Its value ranges from 0 to 1. An R^2 of 0 means that the explanatory variables explain none of the variation of the dependent variable; an R^2 of 1 means that the explanatory variables explain all the variation. The R^2 is .56. This implies that the three explanatory variables explain 56% of the variation in salaries.

What level of R^2 , if any, should lead to a conclusion that the model is satisfactory? Unfortunately, there is no clear-cut answer to this question because the magnitude of R^2 depends on the characteristics of the data being studied and whether the data vary over time or over individuals. Typically, an R^2 is low in cross-sectional studies in which differences in individual behavior are explained. It is likely that these individual differences are caused by many factors that cannot be measured. As a result, the expert cannot hope to explain most of the variation. In time-series studies, in contrast, the expert is explaining the movement of aggregates over time. Because most aggregate time series have substantial

162. The variation is the square of the difference between each Y value and the average Y value, summed over all the Y values.

growth, or trend, in common, it will not be difficult to “explain” one time series using another time series, simply because both are moving together. It follows as a corollary that a high R^2 does not by itself mean that the variables included in the model are the appropriate ones.

Courts should be reluctant to rely solely on a statistic such as R^2 to choose one model over another. Alternative procedures and tests are available.¹⁶³ When utilizing probit or logit models, the R^2 is especially problematic, since (with the dependent variable taking on only two values) predictions will be spread across the 0–1 interval with relatively few values typically being close to either 0 or 1. As a result, R^2 values will be relatively low whether the estimated model coefficients are informative or not. One more instructive measure is the percentage of correct classifications with the prediction being 1 if the predicted value is greater than .5 and 0 otherwise. An alternative measure is a quasi- R -squared measure based on the log likelihood of the estimated model, relative to that of a model with no covariates.

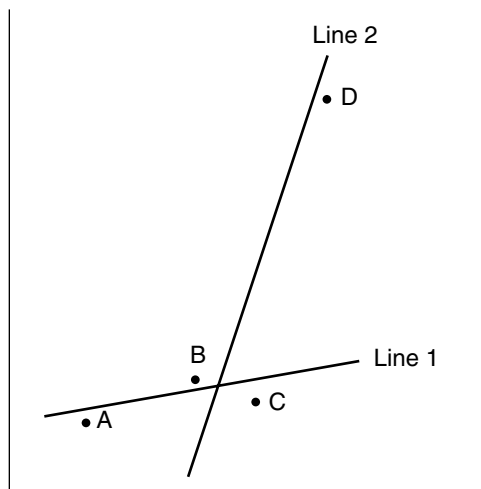
Sensitivity of Least-Squares Regression Results

The least-squares regression line can be sensitive to extreme data points. This sensitivity can be seen most easily in Figure 8. Assume initially that there are only three data points, A, B, and C, relating information about X_1 to the variable Y . The least-squares line describing the best-fitting relationship between Points A, B, and C is represented by Line 1. Point D is called an outlier because it lies far from the regression line that fits the remaining points. When a new, best-fitting least-squares line is re estimated to include Point D, Line 2 is obtained. Figure 8 shows that the outlier, Point D, is an influential data point, because it has a dominant effect on the slope and intercept of the least-squares line. Because least squares minimizes the sum of squared deviations, the sensitivity of the line to individual points sometimes can be substantial.¹⁶⁴

163. These include F -tests and specification error tests. See Pindyck & Rubinfeld, *supra* note 23, at 88–95, 128–36, 194–98.

164. This sensitivity is not always undesirable. In some instances, it may be much more important to predict Point D when a big change occurs than to measure the effects of small changes accurately.

Figure 8. Least-squares regression.



What makes the influential data problem even more difficult is that the effect of an outlier may not be seen readily if deviations are measured from the final regression line. The reason is that the influence of Point D on Line 2 is so substantial that its deviation from the regression line is not necessarily larger than the deviation of any of the remaining points from the regression line.¹⁶⁵ Although they are not as popular as least squares, alternative estimation techniques that are less sensitive to outliers, such as robust estimation, are available. Alternatively, if the sample is relatively large, an expert may choose to remove the outliers before estimating the model.

165. The importance of an outlier also depends on its location in the dataset. Outliers associated with relatively extreme values of explanatory variables are likely to be especially influential. See, e.g., *Fisher v. Vassar College*, 70 F.3d 1420, 1436 (2d Cir. 1995) (court required to include assessment of “service in academic community,” because concept was too amorphous and not a significant factor in tenure review), *rev’d on other grounds*, 114 F.3d 1332 (2d Cir. 1997) (en banc). Conclusory statements of an outlier’s impact are insufficient. See *In re Nektar Therapeutics Sec. Litig.*, 34 F.4th 828, 836 (9th Cir. 2022) (“[t]he complaint does not allege with specificity what the Phase 1 EXCEL results would have been without outlier data”).

A Hypothetical Example

Jane Thompson filed suit in federal court alleging that officials in the police department discriminated against her and a class of other female police officers in violation of Title VII of the Civil Rights Act of 1964, as amended. On behalf of the class, Officer Thompson alleged that she was paid less than male police officers with equivalent skills and experience. Both the plaintiff and the defendant used expert economists with econometric expertise to present statistical evidence to the court in support of their positions.

The plaintiff's expert pointed out that the mean salary of the 40 female officers was \$30,604, whereas the mean salary of the 60 male officers was \$43,077. To show that this difference was statistically significant, the expert put forward a regression of salary (SALARY) on a constant term and a dummy indicator variable (FEM) equal to one for each female and zero for each male. The results were as follows:

$$\begin{array}{lcl} \text{SALARY} = & \$43,077 & - \$12,373 \times \text{FEM} \\ \text{St. Error} & (\$1,528) & (\$2,416) \\ p\text{-value} & <.01 & <.01 \\ R^2 = & .22 & \end{array}$$

The $-\$12,373$ coefficient on the FEM variable measures the mean difference between male and female salaries. Because the standard error is approximately one-fifth of the value of the coefficient, this difference is statistically significant at the 5% (and indeed at the 1%) level. If this is an appropriate regression model (in terms of its implicit characterization of salary determination), one can conclude that it is highly unlikely that the difference in salaries between men and women is due to chance.

The defendant's expert testified that the regression model put forward was the wrong model because it failed to account for the fact that males (on average) had substantially more experience than females. The relatively low R^2 was an indication that there was substantial unexplained variation in the salaries of male and female officers. An examination of data relating to years spent on the job showed that the average male experience was 8.2 years, whereas the average for females was only 3.5 years. The defense expert then presented a regression analysis that added an additional explanatory variable (i.e., a covariate), the years of experience of each police officer (EXP).

The new regression results were as follows:

SALARY = \$28,049 – \$3,860 × FEM + \$1,833 × EXP			
St. Error	(\$2,513)	(\$2,347)	(\$265)
p-value	<.01	<.11	<.01
R ² =	.47		

Experience is itself a statistically significant explanatory variable, with a *p*-value of less than .01. Moreover, the difference between male and female salaries, holding experience constant, is only \$3,860, and this difference is not statistically significant at the 5% level. The defense expert was able to testify on this basis that the court could not rule out alternative explanations for the difference in salaries other than the plaintiff’s claim of discrimination.

The debate did not end here. On rebuttal, the plaintiff’s expert made three distinct points. First, whether \$3,860 was statistically significant or not, it was practically significant, representing a salary difference of more than 10% of the mean female officers’ salaries. Second, although the result was not statistically significant at the 5% level, it was significant at the 11% level.

Third, and most importantly, the expert testified that the regression model was not correctly specified. Further analysis by the expert showed that the value of an additional year of experience was \$2,333 for males on average, but only \$1,521 for females. Based on supporting testimonial experience, the expert testified that one could not rule out the possibility that the mechanism by which the police department discriminated against females was by rewarding males more for their experience than females. The expert made this point clear by running an additional regression in which a further covariate was added to the model. The new variable was an interaction variable, measured as the product of the FEM and EXP variables. The regression results were as follows:

SALARY = \$35,122 – \$5,250 × FEM + \$2,333 × EXP – \$812 × FEM × EXP				
St. Error	(\$2,825)	(\$3,347)	(\$265)	(\$185)
p-value	<.01	<.11	<.01	<.01
R ² =	.65			

The plaintiff’s expert noted that for all males in the sample, FEM = 0, in which case the regression results are given by the equation

$$\text{SALARY} = \$35,122 + \$2,333 \times \text{EXP}$$

However, for females, FEM = 1, in which the corresponding equation is

$$\text{SALARY} = \$29,872 + \$1,521 \times \text{EXP}$$

It appears, therefore, that females are discriminated against not only when hired (i.e., when $EXP = 0$), but also in the reward they get as they accumulate more experience.

The debate between the experts continued, focusing less on the statistical interpretation of any one regression model, but more on the model choice itself, and not simply on statistical significance, but also to practical significance.

Glossary of Terms

alternative hypothesis. See hypothesis test.

association. The degree of statistical dependence between two or more events or variables. Events are said to be associated when they occur more frequently together than one would expect by chance.

bias. Any effect at any stage of investigation or inference tending to produce results that depart systematically from the true values (i.e., the results are either too high or too low). A biased estimator of a parameter differs on average from the true parameter.

coefficient. An estimated regression parameter.

consistent estimator. An estimator that tends to become more and more accurate as the sample size grows.

correlation. A statistical means of measuring the linear association between variables. Two variables are correlated positively if, on average, they move in the same direction; two variables are correlated negatively if, on average, they move in opposite directions.

covariate. A variable that is possibly predictive of an outcome under study; an explanatory variable.

cross-sectional analysis. A type of analysis in which each data point is associated with a different unit of observation (e.g., an individual or a firm) measured at a particular point in time.

degrees of freedom (DF). The number of observations in a sample minus the number of estimated parameters in a regression model. A useful statistic in hypothesis testing.

dependent variable. The variable to be explained or predicted in a multiple regression model.

dummy variable. A variable that takes on only two values, usually 0 and 1, with one value indicating the presence of a characteristic, attribute, or effect (1), and the other value indicating its absence (0).

endogenous variable. A covariate for which there are unobserved factors that affect both the covariate and the dependent variable.

error term. A variable in a multiple regression model that represents the cumulative effect of several sources of modeling error.

estimate. The calculated value of a parameter based on the use of a particular sample.

estimator. The sample statistic that estimates the value of a population parameter (e.g., a regression parameter); its values vary from sample to sample.

ex ante forecast. A prediction about the values of the dependent variable that go beyond the sample; consequently, the forecast must be based on predictions for the values of the explanatory variables in the regression model.

explanatory variable. A variable that is associated with changes in a dependent variable.

ex post forecast. A prediction about the values of the dependent variable made during a period in which all values of the explanatory and dependent variables are known. Ex post forecasts provide a useful means of evaluating the fit of a regression model.

F-test. A statistical test (based on an F -ratio) of the null hypothesis that a group of explanatory variables are jointly equal to 0. When applied to all the explanatory variables in a multiple regression model, the F -test becomes a test of the null hypothesis that R^2 equals 0.

feedback. When changes in an explanatory variable affect the values of the dependent variable, and changes in the dependent variable also affect the explanatory variable. When both effects occur at the same time, the two variables are described as being determined simultaneously.

fitted value. The estimated value for the dependent variable.

fixed-effects model. A regression model that includes a set of dummy variables that account for certain characteristics of individuals in the sample.

heteroscedasticity. When the error associated with a multiple regression model has a nonconstant variance; that is, the error values associated with some observations are typically high, while the values associated with other observations are typically low.

homoscedasticity. When the error associated with a multiple regression model has a constant variance.

hypothesis test. A statement about the parameters in a multiple regression model. The null hypothesis may assert that certain parameters have specified values or ranges; the alternative hypothesis would specify other values or ranges.

independence. When two variables are not correlated with each other (in the population).

independent variable. An explanatory variable that affects the dependent variable but that is not affected by the dependent variable.

influential data point. A data point whose deletion from a regression sample causes one or more estimated regression parameters to change substantially.

instrumental variable. A variable that is both highly correlated with the covariate that is believed to be endogenous and at the same time not directly affected by the dependent variable.

interaction variable. The product of two explanatory variables in a regression model; used in a particular form of nonlinear model.

intercept. The value of the dependent variable when each of the explanatory variables takes on the value of 0 in a regression equation.

least squares (or ordinary least squares). A common method for estimating regression parameters. Least squares minimizes the sum of the squared differences between the actual values of the dependent variable and the values predicted by the regression equation.

linear regression model. A regression model in which the effect of a change in each of the explanatory variables on the dependent variable is the same, no matter what the values of those explanatory variables.

logit model. A discrete dependent-variable regression model in which the estimated coefficients measure the effect of a change in the covariates on the logarithm of the odds that a particular choice will be made or an event will occur.

maximum likelihood estimation. An estimation method that determines values for the parameters of the regression model that maximizes the likelihood that the process described by the model produced the sample data that were observed.

mean; expected value. An average of the outcomes associated with a probability distribution, where the outcomes are weighted by the probability that each will occur.

mean squared error (MSE). The estimated variance of the regression error, calculated as the average of the sum of the squares of the regression residuals.

model. A mathematical representation of an actual situation.

multicollinearity. When two or more variables are highly correlated in a multiple regression analysis. Substantial multicollinearity can cause regression parameters to be estimated imprecisely, as reflected in relatively high standard errors.

multiple regression analysis. A statistical tool for understanding the relationship between two or more variables.

nonlinear regression model. A model having the property that changes in explanatory variables will have differential effects on the dependent variable as the values of the explanatory variables change.

normal distribution. A bell-shaped probability distribution having the property that about 95% of the distribution lies within two standard deviations of the mean.

null hypothesis. In regression analysis, the null hypothesis states that the results observed in a study with respect to a particular variable are no different from what might have occurred by chance, independent of the effect of that variable. See hypothesis test.

one-tailed test. A hypothesis test in which the alternative to the null hypothesis that a parameter is equal to 0 is for the parameter to be either positive or negative, but not both.

outlier. A data point that is more than some appropriate distance from a regression line that is estimated using all the other data points in the sample.

***p*-value.** The significance level in a statistical test; the probability of getting a test statistic as extreme or more extreme than the observed value. The larger the *p*-value, the more likely that the null hypothesis is valid.

parameter. A numerical characteristic of a population or a model.

perfect collinearity. When two or more explanatory variables are correlated perfectly.

population. All the units of interest to the researcher; also, universe.

practical significance. Substantive importance. Statistical significance does not ensure practical significance because, with large samples, small differences can be statistically significant.

probability distribution. The process that generates the values of a random variable. A probability distribution lists all possible outcomes and the probability that each will occur.

probability sampling. A process by which a sample of a population is chosen so that each unit of observation has a known probability of being selected.

probit model. A discrete dependent-variable regression model in which the estimated coefficients measure the effect of a change in a covariate on the probability that a particular choice will be made or an event will occur.

quasi-experiment (or natural experiment). A naturally occurring instance of observable phenomena that yield data that approximate a controlled experiment.

***R*-squared (R^2).** A statistic that measures the percentage of the variation in the dependent variable that is accounted for by all of the explanatory variables in a regression model. *R*-squared is the most used measure of goodness of fit of a regression model.

random-effects regression model. A regression model that views individual-specific constant terms as being randomly distributed across the individual units of observation.

random error term. A term in a regression model that reflects random error (sampling error) that is the result of chance. Consequently, the result obtained

in the sample differs from the result that would be obtained if the entire population were studied.

randomized controlled trial (RCT). A methodology in which the analyst randomly divides potential subjects into two groups: the treatment group, who receive an intervention, and the control group, who do not.

regression coefficient. The estimate of a population parameter obtained from a regression equation that is based on a particular sample; also, regression parameter.

regression residual. The difference between the actual value of a dependent variable and the value predicted by the regression equation.

robust estimation. An alternative to least-squares estimation that is less sensitive to outliers.

robust. When a statistic or procedure does not change much when data or assumptions are slightly modified.

sample. A selection of data chosen for a study; a subset of a population.

sampling error. A measure of the difference between the sample estimate of a parameter and the population parameter.

scatterplot. A graph showing the relationship between two variables in a study; each dot represents one subject. One variable is plotted along the horizontal axis; the other variable is plotted along the vertical axis.

serial correlation. The correlation of the values of regression errors over time.

simultaneity. When a covariate in a regression model affects the dependent variable and is also affected by the dependent variable.

slope. The change in the dependent variable associated with a one-unit change in an explanatory variable in a linear regression model.

spurious correlation. When two variables are correlated, but one is not the cause of the other.

standard deviation. The square root of the variance of a random variable. The variance is a measure of the spread of a probability distribution about its mean; it is calculated as a weighted average of the squares of the deviations of the outcomes of a random variable from its mean.

standard error of forecast (SEF). An estimate of the standard deviation of the forecast error; it is based on forecasts made within a sample in which the values of the explanatory variables are known with certainty.

standard error of the coefficient; standard error (SE). A measure of the variation of a parameter estimate about the true parameter. The standard error is a standard deviation that is calculated from the probability distribution of estimated parameters.

standard error of the regression (SER). An estimate of the standard deviation of the regression error; it is calculated as the square root of the average of the squares of the residuals associated with a particular multiple regression analysis.

statistical significance. A test used to evaluate the degree of association between a dependent variable and one or more explanatory variables. If the calculated p -value is smaller than 5%, the result is said to be statistically significant (at the 5% level). If p is greater than 5%, the result is statistically insignificant (at the 5% level).

t -statistic. A test statistic that describes how far an estimate of a parameter is from its hypothesized value (i.e., given a null hypothesis). If a t -statistic is sufficiently large (in absolute magnitude), an expert can reject the null hypothesis.

t -test. A test of the null hypothesis that a regression parameter takes on a particular value, usually 0. The test is based on the t -statistic.

time-series analysis. A type of multiple regression analysis in which each data point is associated with a particular unit of observation (e.g., an individual or a firm) measured at different points in time.

two-tailed test. A hypothesis test in which the alternative to the null hypothesis that a parameter is equal to 0 is for the parameter to be either positive or negative, or both.

variable. Any attribute, phenomenon, condition, or event that can have two or more values.

variable of interest. The explanatory variable that is the focal point of a particular study or legal issue.

References on Multiple Regression

- Orley Ashenfelter, Theodore Eisenberg & Stewart J. Schwab, *Politics and the Judiciary: The Influence of Judicial Background on Case Outcomes*, 24 J. Legal Stud. 257 (1995).
- Jonathan A. Baker & Daniel L. Rubinfeld, *Empirical Methods in Antitrust: Review and Critique*, 1 Am. L. & Econ. Rev. 386 (1999).
- Gerald V. Barrett & Donna M. Sansonetti, *Issues Concerning the Use of Regression Analysis in Salary Discrimination Cases*, 41 Personnel Psych. 503 (2006).
- Leo Breinman, *Random Forests*, 45 Machine Learning 5 (2001).
- Thomas J. Campbell, *Regression Analysis in Title VII Cases: Minimum Standards, Comparable Worth, and Other Issues Where Law and Statistics Meet*, 36 Stan. L. Rev. 1299 (1984).
- Catherine Connolly, *The Use of Multiple Regression Analysis in Employment Discrimination Cases*, 10 Population Res. & Pol’y Rev. 117 (1991).
- Arthur P. Dempster, *Employment Discrimination and Statistical Science*, 3 Stat. Sci. 149 (1988).
- The Evolving Role of Statistical Assessments as Evidence in the Courts (Stephen E. Fienberg ed., 1989).
- Michael O. Finkelstein, *The Judicial Reception of Multiple Regression Studies in Race and Sex Discrimination Cases*, 80 Colum. L. Rev. 737 (1980).
- Michael O. Finkelstein & Hans Levenbach, *Regression Estimates of Damages in Price-Fixing Cases*, 46 Law & Contemp. Probs. 145.
- Franklin M. Fisher, *Statisticians, Econometricians, and Adversary Proceedings*, 81 J. Am. Stat. Ass’n 277 (1986).
- Franklin M. Fisher, *Multiple Regression in Legal Proceedings*, 80 Colum. L. Rev. 702 (1980).
- Joseph L. Gastwirth, *Methods for Assessing the Sensitivity of Statistical Comparisons Used in Title VII Cases to Omitted Variables*, 33 Jurimetrics J. 19 (1992).
- Peter Kennedy, *A Guide to Econometrics* (6th ed. 2019).
- Note, *Beyond the Prima Facie Case in Employment Discrimination Law: Statistical Proof and Rebuttal*, 89 Harv. L. Rev. 387 (1975).
- Daniel L. Rubinfeld, *Econometrics in the Courtroom*, 85 Colum. L. Rev. 1048 (1985).
- Daniel L. Rubinfeld & Peter O. Steiner, *Quantitative Methods in Antitrust Litigation*, 46 Law & Contemp. Probs. 69 (1983).
- Daniel L. Rubinfeld, *Statistical and Demographic Issues Underlying Voting Rights Cases*, 15 Evaluation Rev. 659 (1991).
- James H. Stock & Mark W. Watson, *Introduction to Econometrics* (4th ed., 2018).

References on Causal Analysis

- Joseph G. Altonji, Todd E. Elder & Christopher R. Taber, *Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools*, 113 J. Pol. Econ. 151 (2005).
- Isaiah Andrews & Timothy B. Armstrong, *Unbiased Instrumental Variables Estimation Under Known First-Stage Sign*, 8 Quantitative Econ. 479 (2017).
- Isaiah Andrews, James H. Stock & Liyang Sun, *Weak Instruments in IV Regression: Theory and Practice*, 11 Ann. Rev. Econ. 727 (2019).
- Joshua D. Angrist, *Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records*, 80 Am. Econ. Rev. 313 (1990).
- Orley Ashenfelter, *Estimating the Effect of Training Programs on Earnings*, 60 Rev. Econ. & Stat. 47 (1978).
- Andrew C. Baker, David F. Larcker & Charles C. Y. Wang, *How Much Should We Trust Staggered Difference-in-Differences Estimates?*, 144 J. Fin. Econ. 370 (2022).
- John Bound, David A. Jaeger & Regina M. Baker, *Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable Is Weak*, 90 J. Am. Stat. Ass'n 443 (1995).
- David Card, *The Causal Effect of Education on Earnings*, in 3 Handbook of Labor Economics, 1801 (Orley Ashenfelter & David Card eds., 1999).
- David Card & Alan B. Krueger, *Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania*, 84 Am. Econ. Rev. 772 (1994).
- Clément de Chaisemartin & Xavier D'Haultfœuille, *Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects*, 110 Am. Econ. R. 2964 (2020).
- David Chan, Matthew Gentzkow & Chuan Yu, *Selection with Variation in Diagnostic Skill*, 137 Q. J. Econ. 729 (2022).
- Maria De Paola & Vincenzo Scoppa, *The Effects of Managerial Turnover: Evidence from Coach Dismissals in Italian Soccer Teams*, 12 J. Sports Econ. 132 (2012).
- Will Dobbie, Jacob Goldin & Crystal S. Yang, *The Effects of Pretrial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges*, 108 Am. Econ. Rev. 201 (2012).
- Floyd v. City of New York, 959 F. Supp. 2d 540 (S.D.N.Y. 2013).
- Andrew Goodman-Bacon, *Difference-in-Differences with Variation in Treatment Timing*, 225 J. Econometrics 254 (2021).
- Justine S. Hastings, *Vertical Relationships and Competition in Retail Gasoline Markets: Empirical Evidence from Contract Changes in Southern California*, 94 Am. Econ. Rev. 317–28 (2004).
- Guido W. Imbens & Joshua D. Angrist, *Identification and Estimation of Local Average Treatment Effects*, 62 Econometrica 467–75 (1994).
- Louis S. Jacobson, Robert J. Lalonde & Daniel G. Sullivan, *Earnings Losses of Displaced Workers*, 83 Am. Econ. Rev. 685–709 (1993).

- Ginger Zhe Jin & Phillip Leslie, *The Effect of Information on Product Quality: Evidence from Restaurant Hygiene Grade Cards*, 118 Q. J. Econ. 409–51 (2003).
- Peter E. Kennedy, *Sinning in the Basement: What are the Rules? The Ten Commandments of Applied Econometrics*, J. Econ. Surveys 578 (2002).
- Edward E. Leamer, *Specification Searches: Ad Hoc Inference with Nonexperimental Data* (1978).
- Steven D. Levitt, *Using Electoral Cycles in Police Hiring to Estimate the Effect of Police on Crime*, 87 Am. Econ. Rev. 270–90 (1997).
- Charles E. Loeffler, *Pre-Imprisonment Employment Drops: Another Instance of the Ashenfelter Dip?*, 108 J. Crim. Law & Criminology 815–38 (2018).
- Winston Lin, *Agnostic Notes on Regression Adjustments to Experimental Data: Reexamining Freedman's Critique*, 7 Annals Applied Stats. 295–318 (2013).
- Justin McCrary, *The Effect of Court-Ordered Hiring Quotas on the Composition and Quality of Police*, 97 Am. Econ. Rev. 318–53 (2007).
- Tracey L. Meares, *The Law and Social Science of Stop and Frisk*, 10 Ann. Rev. L. & Soc. Sci. 335–52 (2014).
- Ashley Miller, *Principal turnover and student achievement*, 6 Econ. Educ. Rev. 60 (2013).
- Emily Oster, *Unobservable Selection and Coefficient Stability: Theory and Evidence*, 37 J. of Bus. & Econ. Stats. 187–204 (2019).
- Peter H. Rossi, Richard A. Berk & Kenneth J. Lenihan, *Money, Work and Crime: Experimental Evidence* (1980).
- Donald B. Rubin, *Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies*, 66 J. Educ. Psych. 688–701 (1974).
- Lawrence W. Sherman & Richard A. Berk, *The Specific Deterrent Effects of Arrest for Domestic Assault*, 49 Am. Socio. Rev. 261–72 (1984).
- James Stock, Jonathan Wright & Mohiro Yugo, *A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments*, 20 J. Bus. & Econ. Stats. 518–29 (2002).
- James H. Stock & Mark W. Watson, *Introduction to Econometrics* (4th ed. 2019).
- Jeffrey M. Wooldridge, *Econometric Analysis of Cross Section and Panel Data* (2001).

Reference Guide on Survey Research

SHARI SEIDMAN DIAMOND, MATTHEW KUGLER,
AND JAMES N. DRUCKMAN

Shari Seidman Diamond, J.D., Ph.D., is the Howard J. Trienens Professor of Law and Professor of Psychology at Northwestern University and a Research Professor at the American Bar Foundation, Chicago.

Matthew Kugler, J.D., Ph.D., is Professor of Law at Northwestern University.

James N. Druckman, Ph.D., is Martin Brewer Anderson Professor of Political Science at the University of Rochester.

CONTENTS

Introduction, 683

Use of Surveys in Court, 685

A Comparison of surveys Evidence and Individual Testimony, 688

Purpose and Design of the Survey, 689

Addressing Relevant Questions, 689

Role of Attorneys in Survey Design and Administration, 690

Skill and Experience of the Experts Who Designed,

Conducted, or Analyzed the Survey, 691

Skill and Experience of the Experts Who Will Testify

About Surveys Conducted by Others, 691

Population Definition and Sampling, 692

The Survey Universe or Population, 692

The Sampling Frame as an Approximation of the Population, 694

The Sample as a Reflection of the Relevant Characteristics
of the Population, 696

Nonresponse as a Potential Source of Bias, 703

Precautions Taken to Ensure That Only Qualified Respondents
Were Included in the Survey, 706

Survey Questions and Structure, 707

Clarity, Precision, and Lack of Bias in the Framing of the
Survey Questions, 707

What Respondents and Interviewers Knew About the
Survey Purpose and Its Sponsorship, 709

Handling Respondents with No Opinions and Reducing
Respondent Guessing, 711

Use of Open-Ended and Closed-Ended Questions, 713

Reference Manual on Scientific Evidence

Use of Probes to Clarify Ambiguous or Incomplete Answers, 717	
Potential Order Effects, 717	
Appropriate Inclusion of Control Groups, Control Questions, or Other Comparisons, 718	
Benefits and Limitations Associated with the Mode of Data Collection Used in the Survey, 725	
In-Person Interviews, 726	
Telephone Interviews, 727	
Mail Questionnaires, 729	
Internet Surveys, 730	
Mixed-Format Surveys, 733	
Surveys Involving Interviewers, 734	
Selection and Training of the Interviewers, 734	
Procedures Used to Ensure and Determine That the Survey Was Administered to Minimize Error and Bias, 735	
Accuracy of Data Entry, 736	
Disclosure and Reporting, 736	
Early Disclosure About the Survey Methodology and Results, 736	
Completeness and Accuracy of All Relevant Information in the Survey Report, 738	
Concluding Observations, 743	
Glossary of Terms, 744	
References on Survey Research, 747	

Introduction

Sample surveys gather responses from a subset of a population to draw inferences about the population as a whole. They are used to describe or enumerate beliefs, attitudes, or behaviors of persons or other social units.¹ Surveys typically are offered in legal proceedings to establish or refute claims about the characteristics of those individuals or social units.² We focus here primarily on sample surveys with individuals reporting about themselves (e.g., their own beliefs) or their organizations (e.g., the company where they are employed). Such surveys must deal not only with issues of population definition, sampling, and measurement common to all surveys, but also with the specialized issues that arise in obtaining information from human respondents.

In principle, a survey can count or measure every member of the relevant population. A survey that does this full count is sometimes called a census. In practice, however, most surveys typically count or measure only a portion of the individuals or other units that the survey intends to describe. In either case, the goal is to provide information on the relevant population. Sample surveys can be carried out using probability or nonprobability sampling techniques. Although probability sampling offers important advantages over nonprobability sampling,³ various forms of nonprobability sampling are in wide use. Thus, in this reference guide, we discuss both probability samples and nonprobability samples, including their strengths and weaknesses for achieving various purposes.

As a method of data collection, surveys have several crucial potential advantages over less systematic approaches.⁴ When a survey is properly designed,

1. Sample surveys conducted by social scientists “consist of (relatively) systematic, (mostly) standardized approaches to collecting information on individuals, households, organizations, or larger organized entities through questioning systematically identified samples.” James D. Wright & Peter V. Marsden, *Survey Research and Social Science: History, Current Practice, and Future Prospects*, in *Handbook of Survey Research* 1, 3 (James D. Wright & Peter V. Marsden eds., 2d ed. 2010).

2. See, e.g., *Sanderson Farms, Inc. v. Tyson Foods, Inc.*, 547 F. Supp. 2d 491 (D. Md. 2008); *SMS Sys. Maint. Servs. v. Digital Equip. Corp.* 188 F.3d 11 22, 22–23 (1st Cir. 1999). For other examples, see *infra* notes 10–26 and accompanying text.

3. See section titled “The Sample as a Reflection of the Relevant Characteristics of the Population” below.

4. This does not mean that surveys can be relied on to address all questions. For example, if survey respondents had been asked in the days before the attacks of 9/11 to predict whether they would volunteer for military service if Washington, D.C., were to be bombed, their answers may not have provided accurate predictions. Although respondents might have willingly answered the question, their assessment of what they would actually do in response to an attack simply may have been inaccurate. Even the option of a “do not know” choice would not have prevented an error in prediction if they believed they could accurately predict what they would do. Thus, although such a survey would have been suitable for assessing the *predictions* of respondents, it might have provided a very inaccurate estimate of what an actual response to the attack would be. If a survey respondent has limited experience (e.g., a child predicting what she will do as an adult) or the hypothetical is far removed from reality or imaginable reality, the survey is unlikely to provide accurate predictions.

executed, and described, it (1) efficiently presents the responses of a group of individuals or other units (e.g., organizations) and (2) permits an assessment of the extent to which the measured responses of the individuals or other units are likely to adequately represent a relevant population of individuals or other units. All questions asked of respondents and all other measuring devices used (e.g., criteria for selecting eligible respondents) can be examined by the court and the opposing party for objectivity, clarity, and relevance, and all answers or other measures obtained can be analyzed for completeness and consistency.

So that the court and the opposing party can closely scrutinize the survey, the party offering the survey as evidence should describe in detail the design, execution, and analysis of the survey, including (1) a description of the population from which the sample was selected, demonstrating that it was a relevant population for the question at hand; (2) a description of how the sample was drawn and an explanation for why that sample design was appropriate; (3) a report on response rate and the ability of the sample to represent the target population; (4) evidence that respondents were attentive and honest in answering the questions on the survey; and (5) an evaluation of any sources of potential bias in respondents' answers or in the ability of the results from the sample to generalize to the relevant population.

The material covered in this reference guide is intended to assist judges in identifying, narrowing, and addressing issues bearing on the adequacy of surveys either offered as evidence or proposed as a method for developing information. Questions about a survey can be (1) raised from the bench during a pretrial proceeding to determine the admissibility of the survey evidence; (2) presented to the contending experts before trial for their joint identification of disputed and undisputed issues; (3) presented to counsel with the expectation that the issues will be addressed during the examination of the experts at trial; or (4) raised in bench trials to help the judge evaluate what weight, if any, the survey should be given.⁵

All sample surveys should address the issues concerning purpose and design (see section titled "Purpose and Design of the Survey" below), population definition and sampling (see section titled "Population Definition and Sampling" below), and disclosure and reporting (see section titled "Disclosure and Reporting" below). All questionnaire and interview surveys raise methodological issues involving survey questions and structure (see section titled "Survey Questions and Structure" below) and confidentiality (see section titled "Disclosure and Reporting" below). Interview surveys introduce additional issues, such as accuracy of data entry (see section titled "Accuracy of Data Entry" below) and interviewer training and qualifications (see

5. Lanham Act cases involving trademark infringement or deceptive advertising frequently require expedited hearings that request injunctive relief, and so judges may need to be more familiar with survey methodology when considering the weight to accord a survey in these cases than when presiding over cases being submitted to a jury. Even in a case being decided by a jury, however, the court must be prepared to evaluate the methodology of the survey evidence in order to rule on admissibility. See *Daubert v. Merrell Dow Pharms., Inc.*, 509 U.S. 579, 589 (1993); Fed. R. Evid. 702.

section titled “Surveys Involving Interviewers” below). And online surveys raise special issues and questions (see section titled “Internet Surveys” below).

Use of Surveys in Court

Sixty years ago, the question of whether surveys were acceptable evidence was unsettled.⁶ Early doubts about the admissibility of surveys centered on their use of sampling and their status as hearsay evidence. Federal Rule of Evidence 703 settled both matters for surveys by redirecting attention to the “validity of the techniques employed.”⁷ The inquiry under Rule 703 focuses on whether facts or data are “of a type reasonably relied upon by experts in the particular field in forming opinions or inferences upon the subject.”⁸

Because the survey method provides an economical and systematic way to gather information and draw inferences about a large number of individuals or other units, surveys are used widely in business, government, administrative settings, and judicial proceedings.⁹ Both federal and state courts have accepted survey evidence on a variety of issues. Our review of cases citing survey evidence over the ten-year period between 2012 and 2022 revealed cases involving legal issues in administrative law, copyright, criminal law, deceptive advertising, discrimination, employment, patent, trademark, and unfair competition, among others.

Some of these cases cited surveys conducted specifically for litigation, and some cited surveys not prepared for litigation. The topics of both litigation and nonlitigation surveys are diverse.¹⁰ Surveys appear even in the most routine of

6. Hans Zeisel, *The Uniqueness of Survey Evidence*, 45 Cornell L.Q. 322, 345 (1960).

7. Fed. R. Evid. 703 advisory committee note. This focus on the adequacy of the methodology used in conducting and analyzing results from a survey is also consistent with the Supreme Court’s discussion of admissible scientific evidence in *Daubert*, 509 U.S. 579; see also *General Elec. Co. v. Joiner*, 522 U.S. 136, 147 (1997).

8. Fed. R. Evid. 703 advisory committee note.

9. Some sample surveys are so well accepted that they may not even be recognized as surveys. For example, some U.S. Census Bureau data are based on sample surveys, including the widely relied-upon American Community Survey, <https://perma.cc/XLY8-R4YE>. Similarly, the Standard Table of Mortality, which is accepted as proof of the average life expectancy of an individual of a particular age and gender, is based on survey data. Surveys conducted by federal agencies are generally of high quality. Their demographic statistics are often used as benchmarks for nongovernmental research.

10. See, e.g., *Kittle-Aikeley v. Claycomb*, 807 F.3d 913, 926 (8th Cir. 2015) (survey to assess the seriousness of the drug abuse problem in schoolchildren to help justify the school’s decision to drug test its community college student body); *Miller v. Bonta*, 542 F. Supp. 3d 1009, 1021 (S.D. Cal. 2021), *vacated and remanded*, No. 21-55608, 2022 WL 3095986 (9th Cir. Aug. 1, 2022) (in a Second Amendment case, survey used to note that many people in California own firearms); *Missouri v. Biden*, 576 F. Supp. 3d 622, 634 (E.D. Mo. 2021) (survey used to predict workers’ response to a vaccine mandate for federal contractors); *United States v. Cloud*, No. 1:19-cr-02032-SMJ-1, 2021 U.S. Dist. LEXIS 235350, at *13 (E.D. Wash. Dec. 8, 2021) (survey used to determine whether a jury pool was sufficiently biased to warrant a change of venue); *Lord & Taylor LLC v. Zim Integrated Shipping Servs., Ltd.*, 108 F. Supp. 3d 197, 227 (S.D.N.Y. 2015) (survey used to note that 79% of

cases. Vocational experts testifying in Social Security Administration disability hearings regularly rely on a variety of surveys conducted by the Bureau of Labor Statistics, specifically the Occupational Requirements Survey,¹¹ the Occupational Employment Survey,¹² and the National Compensation Survey,¹³ to support their opinions on whether a person would be able to find gainful employment given their documented disabilities. One might think that reliance on these surveys is unexceptional, but federal courts have sometimes struggled to determine whether a given expert's use of survey data is scientifically valid.¹⁴

Employment and discrimination law cases sometimes incorporate survey evidence as well. These surveys may be conducted as part of a company's normal business operations and may be relevant to the performance of a particular employee or the feelings of a subgroup of employees.¹⁵ Other times the surveys may be conducted by plaintiffs to provide evidence of wage and working conditions.¹⁶

In cases in which courts set attorneys' fees, judges are charged with determining reasonable fees.¹⁷ Surveys frequently provide courts with evidence of the prevailing market rates.¹⁸

coastal residents thought the impact of Hurricane Sandy was worse than expected, informing whether it should be treated as an Act of God for contract purposes).

11. *Roberta G. v. Kijakazi*, No. CV 20-07796-DFM, 2021 U.S. Dist. LEXIS 179074, at *6-7 (C.D. Cal. Sept. 20, 2021).

12. *Sok v. Kijakazi*, No. 20-cv-489-wmc, 2021 U.S. Dist. LEXIS 183024, at *12 (W.D. Wis. Sept. 24, 2021); *Dawn L.C. v. Comm'r of Soc. Sec.*, No. 3:20-cv-00626-GCS, 2021 U.S. Dist. LEXIS 191829, at *19 (S.D. Ill. Sept. 24, 2021); *Stephanie D. v. Comm'r of Soc. Sec.*, No. 3:20-CV-0768 (ML), 2021 U.S. Dist. LEXIS 168701, at *8 (N.D.N.Y. Sept. 7, 2021).

13. *Missey v. Saul*, No. 4:20-CV-701-ERW, 2021 U.S. Dist. LEXIS 120435, at *14 (E.D. Mo. June 29, 2021).

14. *Alaura v. Colvin*, 797 F.3d 503, 507-08 (7th Cir. 2015) (describing one apparently common practice in apportioning job availability numbers extracted from the Occupational Employment Survey as "preposterous"). See also *Dawn L.C.*, 2021 U.S. Dist. LEXIS 191829, at *17 (describing the subsequent debate among the district courts).

15. *Spivey v. Mohawk ESV, Inc.*, No. 7:19-cv-00670, 2021 U.S. Dist. LEXIS 111905, at *3 (W.D. Va. June 15, 2021) (use of a manager's staff approval ratings in an age discrimination case); *Griffin v. Shelby Residential & Vocational Servs.*, No. 2:18-cv-2665, 2021 U.S. Dist. LEXIS 111866, at *24 (W.D. Tenn. June 15, 2021) ("The Court finds that the 2016 QA surveys are the most probative and objective evidence of Defendant's rationale for termination."); *Milioto-Maruca v. Lauren*, No. 3:11-CV-2120, 2012 U.S. Dist. LEXIS 173945, at *5 (M.D. Pa. Dec. 7, 2012) (work climate survey was conducted at the stores managed by the plaintiff in response to subordinate complaints).

16. *Medlock v. Taco Bell Corp.*, No. 1:07-cv-01314-SAB, 2015 U.S. Dist. LEXIS 165235 (E.D. Cal. Dec. 9, 2015).

17. Courts are directed to calculate a lodestar figure by multiplying the number of hours reasonably expended on the litigation times a reasonable hourly rate and then adjust upward or downward based on particular features of the work involved. *Jones v. George Fox Univ.*, No. 3:19-cv-0005-JR, 2022 U.S. Dist. LEXIS 162869, at *3 (D. Or. Sept. 9, 2022) (citing *Blum v. Stenson*, 465 U.S. 886, 888 (1984)).

18. See, e.g., *Jones*, 2022 U.S. Dist. LEXIS 162869, at *8-9 (a survey of attorneys was conducted by the Portland State University Survey Research Lab for the Oregon State Bar; the fee

Surveys also are common in the areas of trademark and false-advertising law. Survey evidence has been used to establish whether a proposed mark is a generic term,¹⁹ assess whether a mark has secondary meaning or achieved sufficient fame for a dilution claim,²⁰ evaluate whether a defendant's product presents a likelihood of confusion with a plaintiff's mark,²¹ and probe how consumers would have interpreted an allegedly misleading advertisement.²²

Surveys have additionally been conducted in false-advertising cases to assess the value of the deceptive claim to consumers²³ and in patent cases to assess the value of product features.²⁴ The use of conjoint analysis, a relatively new method of making such value attributions for litigation, is discussed below.²⁵

On occasion, courts ruling on the admissibility of scientific claims have examined surveys of scientific experts to assess the extent to which a theory or

award adopted by the court used the 95th percentile rate in light of the attorney's prior experience before admission to the bar).

19. See, e.g., *Snyder's Lance, Inc. v. Frito-Lay N. Am., Inc.*, 542 F. Supp. 3d 371, 397–99, 401–02 (W.D.N.C. 2021) (contrasting the results of two surveys considering “Pretzel Crisps”); *Primary Children's Med. Ctr. Found. v. Scentsy, Inc.*, No. 2:11-cv-1141-TC, 2012 U.S. Dist. LEXIS 86318, at *12 (D. Utah June 20, 2012) (finding important the results of a Teflon survey analyzing perceptions of “Festival of Trees”).

20. *Warner Bros. Ent. v. Glob. Asylum, Inc.*, No. CV 12–9547 PSG (CWx), 2012 U.S. Dist. LEXIS 185695, at *11 (C.D. Cal. Dec. 10, 2012) (“The survey results show[] that nearly 50 percent of respondents associated the term ‘Hobbit’ with” a single source); *MZ Wallace Inc. v. Fuller*, No. 18cv2265(DLC), 2018 U.S. Dist. LEXIS 214754, at *32–35 (S.D.N.Y. Dec. 20, 2018) (finding helpful a study showing a lack of secondary meaning in the plaintiff's alleged mark); *ProFoot, Inc. v. MSD Consumer Care, Inc.*, No. 11–7079, 2012 U.S. Dist. LEXIS 83427, at *25–26 (D.N.J. June 14, 2012) (insufficient awareness to support a claim of fame).

21. *Under Armour, Inc. v. Battle Fashions, Inc.*, No. 5T9-CV-297-BO, 2021 U.S. Dist. LEXIS 114292, at *5–6 (E.D.N.C. June 18, 2021) (admitting a survey evaluating whether the defendant's use of “I can do all things” infringes on the plaintiff's mark); *Nat'l Fin. Partners Corp. v. Paycom Software, Inc.*, No. 14 C 7424, 2015 U.S. Dist. LEXIS 74700, at *32–33 (N.D. Ill. June 10, 2015) (finding probative one of the two Squirt-style studies conducted to assess likelihood of confusion).

22. *Naimi v. Starbucks Corp.*, 798 F. App'x 67, 69 (9th Cir. 2019) (plaintiff's survey sufficient to establish implied representation at the motion to dismiss stage); *Benson v. Newell Brands, Inc.*, No. 19 C 6836, 2021 U.S. Dist. LEXIS 220986, at *19–22 (N.D. Ill. Nov. 16, 2021); *In re Elysium Health-ChromaDex Litig.*, No. 17-cv-7394 (LJL), 2022 U.S. Dist. LEXIS 25063, at *5–36 (S.D.N.Y. Feb. 11, 2022) (contrasting the results of two experts' false advertising studies).

23. *In re Dial Complete Mktg. & Sales Pracs. Litig.*, 320 F.R.D. 326 (D.N.H. 2017) (alleged damages, claiming that the soap did not kill 99.9% of germs, but some smaller percentage).

24. *Apple, Inc. v. Samsung Elecs. Co.*, No. 11-CV-01846-LHK, 2012 U.S. Dist. LEXIS 90877, at *37–38, *42–43 (N.D. Cal. June 29, 2012); *SimpleAir, Inc. v. Google Inc.*, No. 2:14-CV-11, 2015 U.S. Dist. LEXIS 135915, at *6–7 (E.D. Tex. Oct. 5, 2015); *TV Interactive Data Corp. v. Sony Corp.*, 929 F. Supp. 2d 1006, 1020–22 (N.D. Cal. 2013); *Microsoft Corp. v. Motorola, Inc.*, 904 F. Supp. 2d 1109, 1119–20 (W.D. Wash. 2012).

25. See *infra* pp. 722–25.

technique has received widespread acceptance.²⁶ Such surveys can be valuable in assisting the court, but they must reflect the views of a representative group of recognized experts who have responded to questions about the relevant issue.

In addition, survey methodology has been used creatively to assist federal courts in managing mass torts litigation. For instance, faced with the prospect of conducting discovery concerning 10,000 plaintiffs, the plaintiffs and defendants in *Wilhoite v. Olin Corp.*²⁷ jointly drafted a discovery survey that was administered in person by neutral third parties, thus replacing interrogatories and depositions. It resulted in substantial savings in both time and cost. Greater use of this approach would be beneficial to all parties.

A Comparison of Survey Evidence and Individual Testimony

To illustrate the value of a survey, it is useful to compare the information that can be obtained from a competently done survey with the information obtained by other means. A survey is presented by a survey expert, who testifies about the responses of a substantial number of individuals who have been selected according to an explicit sampling plan and asked the same set of questions. In contrast, a party using a nonsurvey method generally identifies several witnesses who testify about their own characteristics, experiences, or impressions. Although the party has no obligation to select these witnesses in any particular way or to report on how they were chosen, the party is not likely to select witnesses whose attitudes or beliefs conflict with the party's interests. The witnesses who testify are aware of the parties involved in the case and have discussed the case with the party before testifying.

Although surveys are not the only means of demonstrating particular facts, presenting the results of a well-done survey through the testimony of an expert is an efficient way to inform the trier of fact about a large group of potential witnesses. In some cases, courts have described surveys as the most direct form of

26. For instance, courts determined that the polygraph test has failed to achieve general acceptance in the scientific community based on the inconsistent reactions revealed in several surveys. See *United States v. Scheffer*, 523 U.S. 303, 309–10 (1998); *United States v. Bishop*, 64 F. Supp. 2d 1149 (D. Utah 1999); *United States v. Varoudakis*, No. 97-10158-RGS, 1998 WL 151238 (D. Mass. Mar. 27, 1998); *State v. Shively*, 999 P.2d 952, *aff'd*, 999 P.2d 259 (Kan. 2000); *Lee v. Martinez*, 96 P.3d 291, 304–06 (N.M. 2004). In contrast, an eyewitness identification expert was permitted to testify about scientific studies of factors affecting the perceptual ability and memory of eyewitnesses based in part on survey evidence showing general acceptance of those findings by experts within the field. See *People v. Williams*, 830 N.Y.S.2d 452, 465–66 (N.Y. Sup. Ct. 2006).

27. *Wilhoite v. Olin Corp.*, No. CV-83-C-5021-NE (N.D. Ala. filed Jan. 11, 1983). The case ultimately settled before trial. See Francis E. McGovern & E. Allan Lind, *The Discovery Survey*, 51 Law & Contemp. Probs. 41, 49 (1988).

evidence that can be offered.²⁸ Indeed, several courts have drawn negative inferences from the absence of a survey, taking the position that failure to undertake a survey may strongly suggest that a properly done survey would not support the plaintiff's position.²⁹

Purpose and Design of the Survey

Addressing Relevant Questions

Key to evaluating any survey is assessing whether it is relevant to the disputed issues. Surveys conducted in the normal course of business and not in anticipation of, or in response to, litigation may have a lesser risk of bias in favor of the interests of the party conducting the survey, but such surveys may ask irrelevant questions, lack important quality controls, or be conducted on inappropriate populations.³⁰ In contrast, surveys conducted for litigation are more likely to be designed to address the legally relevant issues in the case (e.g., to estimate damages in an antitrust suit or to assess consumer confusion in a trademark case), but may have a greater risk of bias since they are typically solicited by one of the parties to aid in the case. Thus, the content and execution of a survey must be scrutinized whether or not the survey was explicitly designed to provide relevant data on the issue before the court.³¹

28. See, e.g., *Morrison Ent. Grp. v. Nintendo of Am.*, 56 F. App'x 782, 785 (9th Cir. 2003); *Monster, Inc. v. Dolby Lab's Licensing Corp.*, 920 F. Supp. 2d 1066, 1072 (N.D. Cal. 2013); *Heaven Hill Distilleries, Inc. v. Log Still Distilling, LLC*, No. 3:21-cv-190-BJB-CHL, 2021 U.S. Dist. LEXIS 240373 (W.D. Ky. Dec. 16, 2021) (survey is the "most persuasive evidence" of consumer recognition).

29. *Ortho Pharm. Corp. v. Cosprophar, Inc.*, 32 F.3d 690, 695 (2d Cir. 1994); *Henri's Food Prods. Co. v. Kraft, Inc.*, 717 F.2d 352, 357–58 (7th Cir. 1983); *Medici Classics Prods. LLC v. Medici Grp. LLC*, 590 F. Supp. 2d 548, 556 (S.D.N.Y. 2008); *Citigroup v. City Holding Co.*, No. 99 Civ. 10115 (RWS), 2003 U.S. Dist. LEXIS 1845 (S.D.N.Y. Feb. 10, 2003); *Chum Ltd. v. Lisowski*, 198 F. Supp. 2d 530 (S.D.N.Y. 2002); *Cairns v. Franklin Mint Co.*, 24 F. Supp. 2d 1013, 1041 (C.D. Cal. 1998) ("[A] plaintiff's failure to conduct a survey, assuming it has the financial resources to do so, may lead to an inference that the results of such a survey would be unfavorable.").

30. In *Craig v. Boren*, 429 U.S. 190 (1976), the state unsuccessfully attempted to use its annual roadside survey of the blood alcohol level, drinking habits, and preferences of drivers to justify prohibiting the sale of 3.2% beer to males under the age of 21 and to females under the age of 18. The data were biased because it was likely that the male would be driving if both the male and female occupants of the car had been drinking. As pointed out in 2 Joseph L. Gastwirth, *Statistical Reasoning in Law and Public Policy: Tort Law, Evidence, and Health* 527 (1988), the roadside survey would have provided more relevant data if all occupants of the cars had been included in the survey (and if the type and amount of alcohol most recently consumed had been requested so that the consumption of 3.2% beer could have been isolated).

31. See *Merisant Co. v. McNeil Nutritionals, LLC*, 242 F.R.D. 315 (E.D. Pa. 2007).

Role of Attorneys in Survey Design and Administration

An early handbook for judges recommended that survey interviews be “conducted independently of the attorneys in the case.”³² Some courts interpreted this to mean that any evidence of attorney participation is objectionable.³³ A better interpretation is that the attorney should not take part in survey implementation or interact with survey participants.³⁴ However, some attorney involvement in the survey design is often necessary to ensure that relevant questions are directed to a relevant population,³⁵ particularly if the survey expert is not an expert in the substantive area of law under dispute. Federal Rule of Civil Procedure 26(4)(b) does not allow an inquiry into the nature of communications between attorneys and experts, and so the role of attorneys in constructing surveys may not always be fully apparent. The key issues for the trier of fact are the design of the survey, the objectivity and relevance of the questions on the survey, the appropriateness of the population used to guide sample selection, and the method of sample selection. These aspects of the survey are visible to the trier of fact and can be judged on their quality, irrespective of who suggested them. In contrast, the survey administration itself, whether online or in an interview, may not be directly visible. Any potential bias is minimized by having interviewers and respondents blind to the purpose and sponsorship of the survey and by excluding attorneys from any part in administering questionnaires, conducting interviews, tabulating results, and interpreting the data.³⁶

32. Judicial Conference of the United States, Handbook of Recommended Procedures for the Trial of Protracted Cases, 25 F.R.D. 351, 429 (1960).

33. See, e.g., *Boehringer Ingelheim G.m.b.H. v. Pharmadyne Lab'ys*, 532 F. Supp. 1040, 1058 (D.N.J. 1980).

34. *Upjohn Co. v. Am. Home Prods. Corp.*, No. 1-95-CV-237, 1996 U.S. Dist. LEXIS 8049, at *42 (W.D. Mich. Apr. 5, 1996) (objection that “counsel reviewed the design of the survey carries little force with this Court because [opposing party] has not identified any flaw in the survey that might be attributed to counsel’s assistance”). For cases in which attorney participation was linked to significant flaws in the survey design or execution, see *Hurt v. Commerce Energy, Inc.*, No. 1:12-CV-00758, 2015 U.S. Dist. LEXIS 10566 (N.D. Ohio Jan. 29, 2015); *Johnson v. Big Lots Stores, Inc.*, No. 04-321, 2008 U.S. Dist. LEXIS 35316, at *52-53 (E.D. La. Apr. 29, 2008); *United States v. Southern Indiana Gas & Electric Co.*, 258 F. Supp. 2d 884, 894 (S.D. Ind. 2003); and *Gibson v. County of Riverside*, 181 F. Supp. 2d 1057, 1069 (C.D. Cal. 2002).

35. See 6 J. Thomas McCarthy, McCarthy on Trademarks and Unfair Competition § 32:166 (5th ed. 2021). See also Jerre B. Swann, *A History of the Evolution of Likelihood of Confusion Methodologies*, 113 Trademark Rep. 723 (2023) (generally on the importance of context in determining survey design methodology).

36. *Gibson*, 181 F. Supp. 2d at 1068.

Skill and Experience of the Experts Who Designed, Conducted, or Analyzed the Survey

Experts prepared to design, conduct, and analyze a survey generally should have graduate training in psychology (especially social, cognitive, or consumer psychology), sociology, political science, marketing, communication sciences, statistics, or a related discipline; that training should include courses in survey research methods, sampling, measurement, interviewing, and statistics. In some cases, professional experience in teaching or conducting and publishing survey research may provide the requisite background. In all cases, the expert must demonstrate an understanding of best practices in survey methodology, including sampling,³⁷ instrument design (questionnaire and interview construction), and statistical analysis.³⁸ Publication in peer-reviewed journals, authored books, fellowship status in professional organizations, faculty appointments, consulting experience, research grants, and membership on scientific advisory panels for government agencies or private foundations are indications of a professional's area and level of expertise. In addition, some surveys involving highly technical subject matter or specific (sub)populations may require experts to have some further specialized knowledge. Under these conditions, the survey expert also should be able to demonstrate sufficient familiarity with the topic and population (or assistance from an individual on the research team with suitable expertise) to design a survey instrument that will communicate clearly with relevant respondents.

Skill and Experience of the Experts Who Will Testify About Surveys Conducted by Others

Parties often call on an expert to testify about a survey conducted by someone else (e.g., by one of the parties to the suit, or by another entity when the survey was not conducted specifically for the case). The secondary expert's role may be to offer support for a survey commissioned by the party who calls the expert, to critique a survey presented by the opposing party, or to introduce findings or conclusions from a survey not conducted in preparation for litigation or by any of the parties to the litigation. The trial court should take into account the exact issue that the expert seeks to testify about and the nature of the expert's field of

37. The one exception is that sampling expertise would be unnecessary if the survey were administered to all members of the relevant population. *See, e.g.,* McGovern & Lind, *supra* note 27.

38. If survey expertise is being provided by several experts, a single expert may have general familiarity but not special expertise in all these areas.

expertise.³⁹ All experts who give opinions about the adequacy and interpretation of a survey not only should have general skills and experience with surveys and be familiar with all of the issues addressed in this reference guide, but also should demonstrate familiarity with the following properties of the survey being discussed:

1. Purpose of the survey;
2. Survey methodology,⁴⁰ including
 - a. the target population,
 - b. the sampling design used in conducting the survey,
 - c. the survey instrument (questionnaire or interview schedule), and
 - d. (for interview surveys) interviewer training and instruction;
3. Results, including rates and patterns of missing data; and
4. Statistical analyses used to interpret the results.

Population Definition and Sampling

The Survey Universe or Population

One of the first steps in designing or evaluating a survey is to identify the target population (or universe).⁴¹ The target population consists of all elements (e.g., individuals) whose characteristics or perceptions the survey is intended to represent. Thus, in trademark litigation, the relevant population in some disputes may include all prospective and past purchasers of the pertinent party's category of

39. For a discussion of the admissibility of expert opinion testimony, see Liesa L. Richter and Daniel J. Capra, *The Admissibility of Expert Testimony*, "Federal Rule of Evidence 702: An Overview" section, in this manual.

40. See *A & M Records, Inc. v. Napster, Inc.*, No. C 99-05183 MHP, 2000 U.S. Dist. LEXIS 20668, at *22-25 (N.D. Cal. Aug. 10, 2000) (holding that expert could not attest credibly that the surveys upon which he relied conformed to accepted survey principles because of his minimal role in overseeing the administration of the survey and limited expert report); *Hr'g Tr.* at 29-30, *Munchkin Inc. v. Playtex Prods., LLC*, No. CV11-503-AHM(RZx) (C.D. Cal. May 1, 2012), ECF No. 285 (excluding survey when the expert "didn't know how [the survey] was administered[.] . . . didn't know how the panel was selected [and] didn't know what the statistical technique was that was used to weigh the survey and provide weight to it"), cited in *Cohen v. Trump*, No. 3:13-cv-2519-GPC-WVG, 2016 U.S. Dist. LEXIS 117059, at *16-17 (S.D. Cal. Aug. 29, 2016).

41. Identification of the proper target population or universe is recognized uniformly as a key element in the development of a survey. See, e.g., *Manual for Complex Litigation*, Fourth, § 11.493 (2004), <https://www.uscourts.gov/file/3228/download> [hereinafter MCL 4th]; see also 3 McCarthy, *supra* note 35, § 32:166; Council of Am. Survey Rsch. Orgs., *Code of Standards and Ethics for Survey Research* § III.A.3 (2011), <https://perma.cc/698Z-CF86> [hereinafter CASRO]. (Note that CASRO merged with the Marketing Research Association to form the Insights Association in 2017.)

goods or services. Similarly, the population for a discovery survey may include all potential plaintiffs or all employees who worked for Company A between two specific dates. In a community survey designed to provide evidence for a motion for a change of venue, the relevant population consists of all jury-eligible citizens in the community in which the trial is to take place.⁴² The definition of the relevant population is crucial because there may be systematic differences in the responses of members of the population and nonmembers. For example, consumers who are prospective purchasers may know more about the product category than consumers who are not considering making a purchase.

The universe must be defined carefully.⁴³ For example, in a survey testing whether the defendant made misleading representations about their digital evidence–related software used by police departments, the appropriate universe consisted of potential purchasers of the software in the law enforcement community. Instead, the sample consisted of respondents who merely used photographs in their law enforcement work and had no influence on purchase decisions involving the relevant software or were even necessarily potential users of such software.⁴⁴ Defects in the universe may lead to misleading results that should reduce the weight that is given to the survey;⁴⁵ a survey should be excluded if it consists of respondents who do not substantially reflect the characteristics of the relevant population.⁴⁶

42. An additional relevant population may consist of jury-eligible citizens in the community where the party would like to see the trial moved. By questioning citizens in more than one community, the survey can test whether moving the trial is likely to reduce the level of animosity toward the party requesting the change of venue. *See* *United States v. Cloud*, No. 1:19-cr-02032-SMJ-1, 2021 U.S. Dist. LEXIS 260839, at *8–9 (E.D. Wash. Dec. 8, 2021) (order granting a motion for intra-district transfer from Yakima based in part on a telephone survey of jury-eligible respondents in Yakima, Richland, and Spokane); *United States v. Haldeman*, 559 F.2d 31, 140, 151 (MacKinnon, J., dissenting), 176–79 (app. A) (D.C. Cir. 1976) (court denied change of venue over the strong objection of Judge MacKinnon, who cited survey evidence that Washington, D.C., residents were substantially more likely to conclude, before trial, that the defendants were guilty); *see also* *People v. Venegas*, 31 Cal. Rptr. 2d 114, 117 (Cal. Ct. App. 1994) (change of venue denied because defendant failed to show that the defendant would face a less hostile jury in a different court).

43. *See* *Merck Eprova AG v. Brookstone Pharms.*, 920 F. Supp. 2d 404, 418 (S.D.N.Y. 2013) (“In determining whether a challenged advertisement is likely to confuse or mislead customers, courts must look to the person to whom the advertisement is addressed.”) (citation omitted).

44. *See, e.g.,* *Kwan Software Eng’g, Inc. v. Foray Techs., LLC*, 110 U.S.P.Q.2d 1637, 1641–42 (N.D. Cal. 2014). *See also* *Warner Bros., Inc. v. Gay Toys, Inc.*, 658 F.2d 76 (2d Cir. 1981) (surveying child users of the product rather than parent purchasers). Children and some other populations create special challenges for researchers. For example, very young children should not be asked about sponsorship or licensing, concepts that are foreign to them. Concepts, as well as wording, should be age appropriate.

45. *See, e.g.,* *Chi. Mercantile Exchange Inc. v. ICE Clear US, Inc.*, No. 18 C 1376, 2020 WL 1905760, at *12–14 (N.D. Ill. Apr. 12, 2020).

46. *See, e.g.,* *Kwan*, 110 U.S.P.Q.2d at 1642.

The Sampling Frame as an Approximation of the Population

The target population consists of all the individuals or units whose responses the researcher would like to describe. The sampling frame is the source (or sources) from which the sample actually is drawn. The surveyor's job generally is easier if a complete list of every eligible member of the population is available (e.g., all plaintiffs in a discovery survey), so that the sampling frame lists all members of the target population. The survey expert should identify how the sampling frame was compiled—that is, the source(s) used to obtain the list of members of the population. Ideally, even if a list of every member of the population is not available, the survey expert will still have access to demographic characteristics of the full population (e.g., sex, race/ethnicity, age). In some cases, this is straightforward, such as when the population consists of all American residents (so the Census's American Community Survey can provide the demographic information). In other situations, population descriptions may be less easily obtained (e.g., the consumers of a particular product, whose demographic characteristics may be identifiable only from the marketing research of the company that produces the product). In practice, the target population often includes some members who cannot be contacted or who cannot be identified in advance. As a result, reasonable compromises are sometimes required in developing the sampling frame. The survey report should contain (1) a description of the target population, (2) a description of the sampling frame from which the sample is drawn, (3) a discussion of the difference between the two, and, importantly, (4) an evaluation of the likely consequences of that difference.

A survey that provides information about a wholly irrelevant population is itself irrelevant.⁴⁷ Courts are likely to exclude such a survey or accord it minimal

47. A survey aimed at assessing how persons in the trade respond to an advertisement should be conducted on a sample of persons in the trade and not on a sample of consumers. *See* *Home Box Off. v. Showtime/The Movie Channel*, 665 F. Supp. 1079, 1083–84 (S.D.N.Y. 1987), *aff'd in part and vacated in part*, 832 F.2d 1311 (2d Cir. 1987); *J & J Snack Food Corp. v. Earthgrains Co.*, 220 F. Supp. 2d 358, 371–72 (D.N.J. 2002); *Parks, LLC v. Tyson Foods, Inc.*, 186 F. Supp. 3d 405, 419–20 (E.D. Pa. 2016) (giving little weight to a survey aimed at consumers of the plaintiff's products when it should have been targeted at users of the defendant's). *But see* *Lon Tai Shing Co., LTD v. Koch & Lowy*, No. 90 Civ. 4464, 1990 U.S. Dist. LEXIS 19123, at *50–57 (S.D.N.Y. Dec. 14, 1990), in which the judge was willing to find likelihood of consumer confusion from a survey of lighting store salespersons questioned by a survey researcher posing as a customer. The court was persuaded that the salespersons who were misstating the source of the lamp, whether consciously or not, must have reasonably believed that the consuming public would be likely to rely on the salespersons' inaccurate statements about the name of the company that manufactured the lamp they were selling.

weight.⁴⁸ More commonly, however, the sampling frame and the target population have some overlap, but the overlap is imperfect. This is called coverage error, and it has two types: the sampling frame may be underinclusive by excluding part of the target population, or it may be overinclusive by including individuals who are not members of the target population. If the coverage is underinclusive, the survey's value depends on the extent to which the excluded population is likely to respond differently from the included population. Thus, a survey of spectators and participants at running events would be sampling a sophisticated subset of those likely to purchase running shoes. Because this subset would probably consist of the consumers most knowledgeable about the trade dress used by companies that sell running shoes, a survey based on this sampling frame would be likely to substantially overrepresent the strength of a particular design as a trademark. Worse still, the extent of that overrepresentation would be unknown and not susceptible to any reasonable estimation.⁴⁹

In some cases, it is difficult to determine whether a sampling frame that omits some members of the population distorts the results of the survey and, if so, the extent and likely direction of the bias. For example, a trademark survey was designed to test the likelihood of confusing an analgesic currently on the market with a new product that was similar in appearance.⁵⁰ The plaintiff's survey included only respondents who had used the plaintiff's analgesic, and the court found that the target population should have included users of other analgesics, "so that the full range of potential customers for whom plaintiff and defendants would compete could be studied."⁵¹ In this instance, it is unclear whether users of the plaintiff's product would be more or less likely to be confused than users of the defendants' product or users of a third analgesic, but the omission of users

48. See *Varner v. Dometic Corp.*, No. 16-22482-CIV-SCOLA/OTAZO-REYES, 2022 U.S. Dist. LEXIS 127353, at *27–30 (S.D. Fla. Feb. 2, 2022) (survey of consumers excluded because gas-absorption refrigerators are not sold directly to consumers, but rather to manufacturers and OEM and RV dealers); see also *In re Fluidmaster, Inc., Water Connector Components Prods. Liab. Litig.*, No. 14-cv-5696, 2017 WL 1196990, at *29 (N.D. Ill. Mar. 21, 2017); *Wells Fargo & Co. v. WhenU.com, Inc.*, 293 F. Supp. 2d 734 (E.D. Mich. 2003).

49. See *Brooks Shoe Mfg. Co. v. Suave Shoe Corp.*, 533 F. Supp. 75, 80 (S.D. Fla. 1981), *aff'd*, 716 F.2d 854 (11th Cir. 1983); see also *Hodgdon Powder Co. v. Alliant Techsystems, Inc.*, 512 F. Supp. 2d 1178, 1181–82 (D. Kan. 2007) (excluding survey on gunpowder brands distributed at plaintiff's promotional booth at a shooting tournament); *Winning Ways, Inc. v. Holloway Sportswear, Inc.*, 913 F. Supp. 1454, 1467 (D. Kan. 1996) (survey flawed in failing to include sporting goods customers who constituted a major portion of customers). But see *Thomas & Betts Corp. v. Panduit Corp.*, 138 F.3d 277, 294–95 (7th Cir. 1998) (survey of store personnel admissible because relevant market included both distributors and ultimate purchasers).

50. See *Am. Home Prods. Corp. v. Barr Lab'ys, Inc.*, 656 F. Supp. 1058 (D.N.J.), *aff'd*, 834 F.2d 368 (3d Cir. 1987).

51. *Am. Home Prods.*, 656 F. Supp. at 1070.

of other analgesics made it impossible to assess the effect of that difference.⁵² If the sampling frame does not include important groups in the target population, the survey cannot provide information on how the unrepresented members of the target population would have responded.⁵³

An overinclusive sampling frame generally presents less of a problem for interpretation than does an underinclusive sampling frame.⁵⁴ If the survey expert can demonstrate that a sufficiently large and representative subset of respondents in the survey was drawn from the appropriate sampling frame, the responses obtained from that subset can be examined, and inferences about the relevant population can be drawn based on that subset.⁵⁵ If the relevant subset cannot be identified, however, an overbroad sampling frame will reduce (or eliminate) the value of the survey.⁵⁶

The Sample as a Reflection of the Relevant Characteristics of the Population

Identification of a survey population must be followed by selection of a sample that accurately represents that population.⁵⁷ The use of probability sampling techniques maximizes both the representativeness of the survey results and the

52. *See also* Craig v. Boren, 429 U.S. 190 (1976).

53. *See, e.g.,* Amstar Corp. v. Domino's Pizza, Inc., 615 F.2d 252, 263–64 (5th Cir. 1980) (court found both plaintiff's and defendant's surveys substantially defective for a systematic failure to include parts of the relevant population); Scott Fetzer Co. v. House of Vacuums, Inc., 381 F.3d 477, 487–88 (5th Cir. 2004) (universe drawn from plaintiff's customer list is underinclusive and customers are likely to differ in their familiarity with plaintiff's marketing and distribution techniques); Hi Ltd. P'Ship v. Winghouse of Fla., Inc., No. 6:03-cv-116-Orl-22JGG, 2004 U.S. Dist. LEXIS 30687, at *25–26 (M.D. Fla. Oct. 5, 2004) (“[T]he failure to include a single female in the survey, when women comprise nearly a third of Hooters’ customer base and perhaps even more of Winghouse’s clientele, reflects a patently flawed methodology striking at the very heart of the survey’s validity.”).

54. *See* Schwab v. Philip Morris USA, Inc., 449 F. Supp. 2d 992, 1135 (E.D.N.Y. 2006) (“Studies evaluating broadly the beliefs of low tar smokers generally are relevant to the beliefs of ‘light’ smokers more specifically.”); Jacobs v. Fareportal, Inc., No. 8:17CV362, 2020 U.S. Dist. LEXIS 211840, at *25–26 (D. Neb. May 29, 2020) (critique about the overinclusiveness of sampling past as well as future purchasers goes to weight rather than admissibility).

55. *See* Nat’l Football League Props., Inc. v. Wichita Falls Sportswear, Inc., 532 F. Supp. 651, 657–58 (W.D. Wash. 1982).

56. *See* Leelanau Wine Cellars, Ltd. v. Black & Red, Inc., 502 F.3d 504, 518 (6th Cir. 2007) (lower court was correct in giving little weight to survey with overbroad universe); Big Dog Motorcycles, L.L.C. v. Big Dog Holdings, Inc., 402 F. Supp. 2d 1312, 1334 (D. Kan. 2005) (universe composed of prospective purchasers of all t-shirts and caps overinclusive for evaluating reactions of buyers likely to purchase merchandise at motorcycle dealerships). *See also* Schieffelin & Co. v. Jack Co. of Boca, Inc., 850 F. Supp. 232, 246 (S.D.N.Y. 1994).

57. MCL 4th, *supra* note 41, § 11.493.

ability to assess the precision of estimates obtained from the survey. Probability samples range from simple random samples to complex multistage sampling designs that use stratification, clustering of population elements into various groupings, or both. In all forms of probability sampling, each element in the relevant population has a known, nonzero probability of being included in the sample.⁵⁸ These sample selection probabilities do not need to be the same for all population elements (i.e., some members may have a higher or lower chance of being selected than others). If the probabilities are unequal, however, compensatory adjustments should be made in the analysis. Failure to adjust by weighting to reflect the known distribution in the population on relevant characteristics may undermine representativeness and warrant exclusion of the survey.⁵⁹

Probability sampling offers two important advantages over nonprobability sampling. First, a probability sample can provide an unbiased estimate that summarizes the responses of all persons in the population from which the sample was drawn; that is, the expected value of the sample estimate is the population value being estimated. For instance, if a probability sample leads to an estimate that 80% of respondents were confused as to whether an analgesic was an existing or new option, the researcher could be confident that approximately 80% of those in the population would have that belief (within some margin of error). A probability sample can provide an unbiased estimate of the population even when the size of the sample is relatively small. The advantage of larger samples is that they provide more precision (smaller margins of error). In general, the absolute size of the sample, rather than its size relative to the population, determines the precision of the estimate.⁶⁰

Second, probability sampling allows the researcher to calculate a confidence interval that explicitly provides information on the reliability of the sample estimate of the population value (i.e., the value one would obtain if everyone in the population responded). The difference between the estimate and the exact value

58. See Thomas Piazza, *Fundamentals of Applied Sampling*, in *Handbook of Survey Research*, *supra* note 1, at 139, 145.

59. *Fish v. Kobach*, 309 F. Supp. 3d 1048, 1059–61 (D. Kan. 2018), *aff'd sub nom.* *Fish v. Schwab*, No. 18–3133, 2020 U.S. App. LEXIS 13723 (10th Cir. Apr. 29, 2020) (expert's survey intended to represent the eligible voting population of Kansas was excluded, in part, for failure to weight responses to reflect the distribution of educational attainment and household income level in the population).

60. An exception to this rule applies when a population is (identifiably) finite and the sample size is large relative to the population size, typically taken to occur when the sample size is greater than 5% of the total population (e.g., the population strictly includes 1,000 people, and the sample includes more than 50 people). In this case, it is appropriate to adjust the standard error/margin of error by the finite population correction (FPC) factor. The FPC factor incorporates both the size of the sample and its share of the population in determining the degree of precision. See William G. Cochran, *Sampling Techniques* 24–25 (3d ed. 1977).

is called the sampling error.⁶¹ Thus, suppose a survey collected responses from a simple random sample of 400 dentists selected from the population of all dentists licensed to practice in the United States and found that 20% of them mistakenly believed that a new toothpaste, Goldgate, was manufactured by the makers of Colgate. A survey expert could properly compute a confidence interval around the 20% estimate obtained from this sample. If the survey were repeated a large number of times, and a 95% confidence interval was computed each time, 95% of the confidence intervals would include the actual percentage of dentists in the entire population who would believe that Goldgate was manufactured by the makers of Colgate.⁶² In this example, the margin of error is $\pm 4\%$, and so the confidence interval is the range between 16% and 24%—that is, the estimate (20%) plus or minus 4%.

All sample surveys (i.e., surveys that do not measure responses from every member of the population) produce estimates of population values, not exact measures of those values. Assuming a probability sample, a confidence interval describes how stable the mean response in the sample is likely to be. The width of the confidence interval depends on three primary characteristics:

1. Size of the sample (larger samples produce narrower intervals);
2. Variability of the response being measured (more variability leads to wider intervals); and
3. Confidence level the researcher wants to have.⁶³

Traditionally, scientists adopt the 95% level of confidence, which means that if one hundred samples of the same size were drawn, the confidence interval expected for ninety-five of the samples would be expected to include the true population value.⁶⁴

Stratified probability sampling uses what is known about the population characteristics to correct for imbalances produced by random sampling. Consider the dentist example presented earlier: if it is known that 60% of dentists have at least

61. See the glossary of this chapter, and David H. Kaye & Hal S. Stern, *Reference Guide on Statistics and Research Methods*, in this manual, for a more detailed definition of sampling error.

62. Actually, because survey interviewers would be unable to locate some dentists, and some dentists would be unwilling to participate in the survey, technically the population to which this sample would be projectable would be all dentists with current addresses who would be willing to participate in the survey if they were asked. The expert should be prepared to discuss possible sources of bias due to, for example, an address list that is not current.

63. When the sample design does not use a simple random sample, the confidence interval will be affected.

64. To increase the likelihood that the confidence interval contains the actual population value (e.g., from 95% to 99%) without increasing the sample size, the width of the confidence interval can be expanded. An increase in the confidence interval brings an increase in the confidence level. For further discussion of confidence intervals, see the glossary to the current chapter, and David H. Kaye & Hal S. Stern, *Reference Guide on Statistics and Research Methods*, in this manual.

fifteen years of experience and 40% have fewer years, the researcher could randomly sample dentists within each of those separate categories to obtain a sample that reflected those proportions. This would ensure the sample was unbiased with respect to years of experience—that proportion would be set by the researcher—and would therefore reduce sampling error.⁶⁵ Disproportionate sampling from subgroups may be used to enable the survey to provide separate estimates for particular subgroups but should be weighted (as discussed below) to reflect the population as a whole when conducting an overall analysis.

The last decade has seen a dramatic rise in the availability of nonprobability samples. Most of these samples come in one of three general varieties. The first variety is drawn from large nonprobability internet panel samples overseen by a vendor. For these panels, individuals opt in (e.g., via advertisements) to receive compensation for taking surveys. The companies that oversee these panels can produce a set of respondents that consists of members of the target population (e.g., the population of elementary school teachers) and matches the distribution of that population on specified characteristics (e.g., percentage of women and minorities). These samples vary in their ability to hit benchmarks, but some perform quite well.⁶⁶ Regardless, these samples differ from probability samples because they are not drawn from a list of the entire population. Instead, there is an attempt to identify relevant respondents and to balance the sample to resemble the population on certain observable features (e.g., the sample and population have the same distribution of gender, age, income).

The second source of nonprobability samples is crowdsourcing labor market platforms, the best-known of which is Amazon's Mechanical Turk (MTurk). Researchers can directly use MTurk or analogous platforms to hire individuals to complete tasks, including taking surveys, for direct compensation. Here, researchers can try to draw samples that match populations on observable variables, but it is often difficult since these platforms do not easily allow for the kind of selective participant invitations that permit managed panels to build custom samples.

The third source of nonprobability samples are purposive samples. These use nonprobability methods designed to target hard-to-reach populations (for whom probability sampling is extremely difficult or impossible), such as low-income people, young people, Indigenous people, and people in poor health.⁶⁷ These sampling methods are sometimes the subject of judicial concerns about researcher bias in selecting respondents, as these methods require the researcher to engage

65. See *Pharmacia Corp. v. Alcon Lab'ys, Inc.*, 201 F. Supp. 2d 335, 365 (D.N.J. 2002).

66. Lynn Vavreck & Douglas Rivers, *The 2006 Cooperative Congressional Election Study*, 18 J. Elections, Pub. Op., & Parties, 355–66 (2008), <https://doi.org/10.1080/17457280802305177>.

67. Hard-to-Survey Populations (Roger Tourangeau et al. eds., 2014). Abdolreza Shaghghi et al., *Approaches to Recruiting 'Hard-To-Reach' Populations into Research: A Review of the Literature*, 1 Health Promotion Persps. 86, 94 (2011), <https://doi.org/10.5681/hpp.2011.009>.

in active and targeted outreach.⁶⁸ They also require particular attention to language and survey mode.⁶⁹

Generally speaking, researchers employ nonprobability samples because they are substantially less expensive, often allow for larger samples (due to the lower cost), may include a higher number of individuals from less prevalent subgroups (e.g., members of racial or ethnic minority groups) that could be of interest, and/or offer the only option for hard-to-reach populations. Indeed, the use of nonprobability samples has been common in litigation for some time. For instance, an overwhelming majority of the consumer surveys conducted for Lanham Act litigation present results from nonprobability samples,⁷⁰ and the standard modern practice is for these surveys to be conducted online using professionally managed nonprobability internet panels (i.e., the first type discussed above).⁷¹ A common rationale for admitting such surveys into evidence is that they are used widely in marketing research and that “results of these studies are used by major American companies in making decisions of considerable consequence.”⁷²

Nonprobability samples are appropriate for some types of litigation surveys but still carry important limitations. In many cases, researchers cannot compute response rates for nonprobability samples because they are unaware of how many potential respondents viewed the invitation to participate in the survey and decided not to do so (e.g., researchers often just post participation invitations to those in a labor market without knowing how many people viewed it). In contrast, when trying to assess the base rate of some variable in the national population, probability samples offer superior accuracy relative to nonprobability samples

68. See, e.g., *Branson v. All. Coal, LLC*, No. 4:19-CV-00155-JHM-HBB, 2022 U.S. Dist. LEXIS 123731, at *23–24 (W.D. Ky. July 13, 2022) (rejecting the use of purposive sampling in a class certification survey: “Put simply, the inclusion of purposefully selected data questions whether the data accurately represents the population as a whole. To preserve the reliability of the discovery data, no party shall select certain opt-in plaintiffs from whom to collect information or depose.”).

69. Robin Bayes et al., *Studying Science Inequities: How to Use Surveys to Study Diverse Populations*, 700 *Annals Am. Acad. Pol. & Soc. Sci.*, 220, 233 (2022), <https://doi.org/10.1177/00027162221093970>.

70. Jacob Jacoby & Amy H. Handlin, *Non-Probability Sampling Designs for Litigation Surveys*, 81 *Trademark Rep.* 169, 173 (1991). For probability surveys conducted in trademark cases, see *James Burrough, Ltd. v. Sign of Beefeater, Inc.*, 540 F.2d 266 (7th Cir. 1976); *Nightlight Systems v. Nitelites Franchise Systems*, No. 1:04-CV-2112-CAP, 2007 U.S. Dist. LEXIS 95565 (N.D. Ga. July 17, 2007); *National Football League Properties, Inc. v. Wichita Falls Sportswear, Inc.*, 532 F. Supp. 651 (W.D. Wash. 1982).

71. Matthew B. Kugler & R. Charles Henn, *Internet Surveys in Trademark Cases: Benefits, Challenges, and Solutions*, in *Trademark and Deceptive Advertising Surveys* 291, 293 (Shari Diamond & Jerre Swann eds., 2d ed. 2022).

72. *Nat’l Football League Props., Inc. v. N.J. Giants, Inc.*, 637 F. Supp. 507, 515 (D.N.J. 1986). A survey of members of the Council of American Survey Research Organizations, the national trade association for commercial survey research firms in the United States, revealed that 95% of the in-person independent contacts in studies done in 1985 took place in malls or shopping centers. Jacoby & Handlin, *supra* note 70, at 172–73, 176.

on validated demographic measures such as sex, age, race, and ethnicity,⁷³ as well as validated behavioral measures such as vaccine uptake.⁷⁴ This is because probability samples are more representative of the underlying population. In evaluating the representativeness of a nonprobability sample relative to the population, the expert should consider both observable and nonobservable sample demographics. On critical variables, it may be necessary to include the joint distributions of the measured variables (e.g., the percentage of minority women).⁷⁵ More generally, there is always the possibility of unobserved relevant factors even in a well-conducted nonprobability survey.⁷⁶ For instance, a nonprobability sample with appropriate quotas may still not include typical members from each quota group. This can lead to inaccuracies in point estimates. Nonprobability samples can provide unbiased estimates if the sample is representative on all variables that correlate with the outcome of interest, or has been weighted to be representative, but this can be challenging when it comes to unobserved factors.⁷⁷

Ultimately, the importance of using a probability sample depends on the survey's goal. If the survey is designed to make a causal inference based on differences between randomly assigned experimental conditions (40% of people in

73. Bo MacInnis et al., *The Accuracy of Measurements with Probability and Nonprobability Survey Samples: Replication and Extension*, 82 Pub. Op. Q. 707 (2018), <https://doi.org/10.1093/poq/nfy038>. The authors compare the accuracy of probability surveys (RDD telephone and internet), internet surveys that combine nonprobability and probability samples, and internet surveys that are fully nonprobability (but with different types of compensation). This study is the most extensive of its kind, using a set of fifty measures and forty benchmark variables from federal face-to-face surveys with high response rates. The analyses showed that probability samples provide more accurate estimates of population quantities than nonprobability samples as well as samples that combined methods.

74. Valerie C. Bradley et al., *Unrepresentative Big Surveys Significantly Overestimated US Vaccine Uptake*, 600 Nature 695 (2021), <https://doi.org/10.1038/s41586-021-04198-4>.

75. Connor Huff & Dustin Tingley, "Who Are These People?" *Evaluating the Demographic Characteristics and Political Preferences of MTurk Survey Respondents*, 2 Rsch. & Politics 1 (2015), <https://doi.org/10.1177/2053168015604648>.

76. The court in *Kinetic Concepts, Inc. v. BlueSky Medical Corp.*, No. SA-03-CA-0832, 2006 U.S. Dist. LEXIS 60187, at *14–17 (W.D. Tex. Aug. 11, 2006), found the plaintiff's survey using a nonprobability sample to be admissible and permitted the plaintiff's expert to present results from a survey using a convenience sample. The court then assisted the jury by providing an instruction on the differences between probability and convenience samples and the estimates obtained from each.

77. See, e.g., Robert M. Groves et al., *Survey Methodology* (2d ed. 2009). Groves and colleagues point out that using statistical significance tests and confidence intervals with nonprobability samples is technically inappropriate since it becomes impossible to obtain an unbiased estimate of sampling variance. That said, they accept that nonprobability samples can approximate probability samples when the researcher knows what population attributes correlate with the key statistics and ensures that the sample is balanced on those attributes (i.e., using quotas). *Id.* at 409–10. Vasia Vehovar and colleagues agree: "We thus recommend to be more openly accepting of the reality of using a standard statistical inference approach as an approximation in non-probability settings" as long as the nature of how the sample is drawn is clear. Vasia Vehovar et al., *Non-Probability Sampling*, in *The SAGE Handbook of Survey Methodology* (2016), <https://doi.org/10.4135/9781473957893>.

the experimental condition were confused by defendant's advertising, but only 20% by control advertising), a survey-experiment using a nonprobability sample of relevant respondents will generally suffice. If the aim is to obtain a precise measure of a particular feature of a population, however, a nonprobability sample faces challenges.⁷⁸ An expert should consider the potential biases in the sample and whether they impact factors correlated with the variable being estimated, to assess the likely magnitude of the biases (i.e., how much they are likely to affect the estimates). The expert should be prepared to explain how observable and unobservable sample characteristics might impact their estimate.

When using a nonprobability sample in an experimental context (i.e., a survey-experiment), the primary concern about the characteristics of the sample is whether they modify the causal relationship being tested in the survey.⁷⁹ For instance, if more and less experienced dentists react the same way to the name Goldgate for toothpaste—if experience does not moderate (that is, change) the effect of the name—a survey-experiment that does not include more experienced dentists will still produce results that generalize across experience level. In contrast, if reaction to the name does vary with experience—if experience *does* moderate the reaction—a sample that includes only less experienced dentists will produce a biased result. The survey expert should identify any potential sample characteristics that may moderate the effects being assessed and demonstrate that the sampling approach takes them into account.

In sum, the historic gold standard was a probability sample because it could provide unbiased estimates of the population. Yet, increased costs of probability samples, more ways to obtain nonprobability samples, challenges with hard-to-reach populations, and nonresponse bias (discussed below) have all led to greater use of nonprobability samples. When such samples are used, it is vital to clarify how the sample compares to known-probability benchmarks (e.g., demographic

78. However, even if the initial sampling approach uses a probability sample, systematic non-response can undermine the representativeness of the sample and hence the accuracy of a precise estimate. This effect has been demonstrated by studies of the effect of a two-stage process to recruit survey participants in which the initial invitation used probability sampling through U.S. Postal Service mailings, phone contact, and modest incentives, and a follow-up effort involving FedEx mailings, in-person recruitment by field interviewers, and enhanced incentives attempted to recruit initial nonresponders. Comparisons of the two groups indicated small but statistically significant differences in reported political views and income level. The importance of such differences will depend on the claim being made for the accuracy of the estimate. Ipek Bilgen et al., *The Undercounted: Measuring the Impact of 'Nonresponse Follow-up' on Research Data* (2019), <https://perma.cc/P5RS-268Q>.

79. See James N. Druckman & Cindy D. Kam, *Students as Experimental Participants: A Defense of the 'Narrow Data Base'*, in *Cambridge Handbook of Experimental Political Science* (2011); James N. Druckman, *Experimental Thinking: A Primer on Social Science Experiments* (2022).

characteristics and other features that could influence the outcome being studied), what adjustments may have been made (e.g., weights; see below) and how biases may influence inferences—particularly when the goal is to arrive at a precise estimate of a population value. Nonprobability samples tend to be on stronger footing when used in an experimental context, but even here it is vital for the researcher to discuss potential variables that could moderate the experimental effect and whether the sampling approach influenced the impact of those moderators.

Nonresponse as a Potential Source of Bias

Even when a sample is drawn randomly from a complete list of elements in the target population (i.e., a probability sample), responses or measures are typically obtained from only part of the sample. If this lack of response is distributed randomly, valid inferences about the population can be drawn with assurance using the measures obtained from the available sample. But nonresponse often is not random. For example, persons who are single typically have three times the “not at home” rate in U.S. Census Bureau surveys as do people living with family.⁸⁰ Nonresponse also occurs when contact with the sampled individual is made, but that person declines to participate. This problem arose recently in political polls where Republicans seemed less likely to participate than Democrats.⁸¹ Efforts to increase response rates include making several attempts to contact potential respondents, sending advance letters,⁸² and providing financial or nonmonetary incentives for participating in the survey.⁸³

The key to evaluating the effect of nonresponse in a survey is to determine the extent to which nonrespondents differ from respondents in a way that would impact the results of the survey. If nonresponse has biased the pattern of responses, the size and direction of that bias need to be assessed. On some occasions, it may be possible to anticipate systematic patterns of nonresponse. For example, high-volume medical professionals may be less willing to respond to a survey than those with lower-volume practices. If the survey includes questions about volume of

80. 2 Gastwirth, *supra* note 30, at 501.

81. Courtney Kennedy et al., Pew Research Center, *Confronting 2016 and 2020 Polling Limitations* (2021), <https://perma.cc/P5CC-HJH2>; see also Joshua D. Clinton et al., *Reluctant Republicans, Eager Democrats? Partisan Nonresponse and the Accuracy of 2020 Presidential Pre-election Telephone Polls*, 86 Pub. Op. Q. 247 (2022), <https://doi.org/10.1093/poq/nfac011>.

82. Edith De Leeuw et al., *The Influence of Advance Letters on Response in Telephone Surveys: A Meta-Analysis*, 71 Pub. Op. Q. 413 (2007), <https://doi.org/10.1093/poq/nfm014> (advance letters effective in increasing response rates in telephone as well as mail and face-to-face surveys).

83. Erica Ryu et al., *Survey Incentives: Cash vs. In-Kind; Face-to-Face vs. Mail; Response Rate vs. Nonresponse Error*, 18 Int'l J. Pub. Op. Rsch. 89 (2006), <https://doi.org/10.1093/ijpor/edh089>.

practice, the expert can assess how experience level may have affected the pattern of results.⁸⁴

Although high response rates are desirable because they reduce the potential impact of nonresponse bias, they are increasingly difficult to achieve. Survey nonresponse rates have risen substantially in the twenty-first century, along with the costs of obtaining responses, and so the issue of nonresponse has attracted considerable attention from survey researchers.⁸⁵

Researchers have developed a variety of approaches to adjust for nonresponse, including weighting obtained responses in proportion to known demographic characteristics of the target population (which we discuss below), comparing the pattern of responses from early and late responders to mail surveys or the pattern of responses from easy-to-reach and hard-to-reach responders in telephone surveys, and imputing estimated responses to nonrespondents based on known characteristics of those who have responded. All of these techniques can only approximate the response patterns that would have been obtained if nonrespondents had responded. Nonetheless, they are useful for testing the robustness of the estimates obtained from responders.

To assess the general impact of the lower response rates, researchers have compared results obtained from surveys with varying response rates.⁸⁶ Surprisingly comparable results have been obtained in many surveys with varying response rates, suggesting that surveys may achieve reasonable estimates even with relatively low response rates. The key is whether nonresponse is associated with systematic differences in response that cannot be adequately modeled or assessed. Generally, the representativeness of a sample (regardless of response rate) matters much more for accurate inference than the response rate.⁸⁷ Determining whether

84. In *People v. Williams*, 830 N.Y.S.2d 452 (N.Y. Sup. Ct. 2006), a published survey of experts in eyewitness research was used to show general acceptance of various eyewitness phenomena. See Saul Kassin et al., *On the "General Acceptance" of Eyewitness Testimony Research: A New Survey of the Experts*, 56 Am. Psychologist 405 (2001), <https://doi.org/10.1037/0003-066X.56.5.405>. The survey included questions on the publication activity of respondents and compared the responses of those with high and low research productivity. Productivity levels in the respondent sample suggested that respondents constituted a blue-ribbon group of leading researchers. *Williams*, 830 N.Y.S.2d at 457, 459 n.26. See also *Pharmacia Corp. v. Alcon Lab'ys, Inc.*, 201 F. Supp. 2d 335 (D.N.J. 2002).

85. E.g., Richard Curtin et al., *Changes in Telephone Survey Nonresponse over the Past Quarter Century*, 69 Pub. Op. Q. 87 (2005), <https://doi.org/10.1093/poq/nfn002>. See also Robert M. Groves & Emilia Peytcheva, *The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis*, 72 Pub. Op. Q. 167 (2008), <https://doi.org/10.1093/poq/nfn011>; Peter Miller et al., Federal Committee on Statistical Methodology, *A Systematic Review of Nonresponse Bias Studies in Federally Sponsored Surveys* (2020), <https://perma.cc/6WCQ-AFNK>.

86. E.g., Daniel M. Merkle & Murray Edelman, *Nonresponse in Exit Polls: A Comprehensive Analysis*, in *Survey Nonresponse* 243–57 (Robert M. Groves et al. eds., 2002) (finding minimal nonresponse error associated with refusals to participate in in-person exit polls); see also Jon A. Krosnick, *Survey Research*, 50 Ann. Rev. Psych. 537 (1999), <https://doi.org/10.1146/annurev.psych.50.1.537>.

87. Bo MacInnis et al., *The Accuracy of Measurements with Probability and Nonprobability Survey Samples: Replication and Extension*, 82 Pub. Op. Q. 707 (2018), <https://doi.org/10.1093/poq/nfy038>;

the level of nonresponse in a survey seriously impairs inferences drawn from the results generally requires an analysis of the determinants of nonresponse. For example, even a survey with a high response rate may seriously underrepresent some portions of the population, such as the unemployed or the poor. The survey expert should be prepared to provide evidence on the potential impact of nonresponse on the survey results.

In surveys that include sensitive or difficult questions, some respondents may refuse to provide answers or may provide incomplete answers.⁸⁸ To assess the impact of nonresponse to a particular question, the survey expert should analyze the differences between those who answered and those who did not answer. Procedures to address the problem of missing data include recontacting respondents to obtain the missing answers and using a respondent's other answers to predict the missing response (i.e., imputation).⁸⁹

Survey researchers commonly use weights, even with many probability samples, to address coverage gaps and control for differential nonresponse among subgroups. Weighting adjustments can sometimes improve the accuracy of survey results.⁹⁰ Responses are weighted to reflect distributions in the target population, typically demographics.⁹¹ When the population includes all Americans, the U.S. Census or American Community Survey can provide demographic population figures that can guide how to weight the survey responses. For instance, if the population consists of 50% men but the sample contains only 40% men, then male sample respondents will be weighted to count more in computations from the sample (and women will be counted less) to better approximate the target

Scott Keeter, *Evidence About the Accuracy of Surveys in the Face of Declining Response Rates*, in *The Palgrave Handbook of Survey Research* 19–22 (David L. Vannette & Jon A. Krosnick eds., 2018), https://doi.org/10.1007/978-3-319-54395-6_4.

88. See Roger Tourangeau et al., *The Psychology of Survey Response* (2000); *California v. Ross*, 358 F. Supp. 3d 965 (N.D. Cal. 2019) (surveys were used to show that if a question on citizenship was included in the 2020 census, some individuals would be less likely to participate or answer the citizenship question, and that this refusal was likely to be higher in the Latinx community).

89. See Paul D. Allison, *Missing Data*, in *Handbook of Survey Research*, *supra* note 1, at 630; see also Groves & Peytcheva, *supra* note 85.

90. Heidi Jensen et al., *The Impact of Non-Response Weighting in Health Surveys for Estimates on Primary Health Care Utilization*, 32 *Eur. J. Public Health* 450 (2022), <https://doi.org/10.1093/eurpub/ckac032>. But see Kenneth Bollen et al., *Are Survey Weights Needed? A Review of Diagnostic Tests in Regression Analysis*, 3 *Ann. Rev. Stat. Application* 375 (2016), <https://doi.org/10.1146/annurev-statistics-011516-012958> (suggesting that weighting is sometimes unhelpful and inefficient).

91. Jelke Bethlehem & Mario Callegaro, *Part IV Weighting Adjustments*, in *Online Panel Research: A Data Quality Perspective* 264–72 (Mario Callegaro et al. eds., 2014). See also *Fish v. Kobach*, 309 F. Supp. 3d 1048, 1089 (D. Kan. 2018) (criticizing an expert for not assigning weights to their results given the differences between their sample's demographics and the expected population demographics); *California v. Ross*, 358 F. Supp. 3d 965, 984 (N.D. Cal. 2019) (commenting that the expert used weighting to balance the demographics of the sample).

population.⁹² When responses are weighted, it is crucial to be transparent about how the data were weighted, and to acknowledge that weighting reduces statistical power and, hence, the precision of the estimates.⁹³

There are three main challenges with survey weights. First, the survey expert needs to know the true proportions in the population (how many are men, aged 18–35, college educated, etc.), or the sample weights will be largely guesses. Weighting cannot compensate for lack of surveyor knowledge about the underlying distributions in the population. Second, weighting does not actually increase sample size. If a sample of forty African American respondents is weighted at 1.5 (meaning each counts as one and a half people for purposes of analysis), the size of their subsample (and consequently the number used to calculate the margin of error of that subsample) is still forty, not sixty. In fact, weighting decreases the precision of estimates, leading to larger confidence intervals. Finally, weighting cannot correct for unobserved factors. If the survey has recruited unrepresentative members of a subpopulation (for example, atypical Republicans), weighting cannot somehow make them representative. This may be a particular problem when weighting is used to substantially amplify the responses of a small number of people.⁹⁴

Precautions Taken to Ensure That Only Qualified Respondents Were Included in the Survey

In a carefully executed survey, each potential respondent is questioned on the attributes that determine their eligibility to participate in the survey. Thus, the initial questions screen potential respondents to determine if they are members of the target population of the survey (e.g., Does she own a dog? Does he live within ten miles of where he works?). The screening questions must be drafted so that they do not appeal to or deter specific groups within the target population or convey information that will influence a respondent's answers on the main survey (i.e., create a context effect). For example, if respondents must be prospective or recent purchasers of Sunshine orange juice in a trademark survey

92. The process often involves iteratively adjusting for multiple demographics. The idea is to assign an adjustment weight to each respondent such that those from underrepresented groups receive weights greater than 1 and those from overrepresented groups receive weights less than 1. There are a host of ways to weight, and caps should be placed on how much a given observation can be weighted so that one respondent does not have a grossly disproportionate effect on the outcomes. See Matthew DeBell, *Computation of Survey Weights*, in *The Palgrave Handbook of Survey Research* 519–27 (2018).

93. Thomas Piazza, *Fundamentals of Applied Sampling*, in *Handbook of Survey Research*, *supra* note 1. See also *Texas v. Holder*, 888 F. Supp. 2d 113, 131 (D.D.C. 2012) (criticizing an expert for inconsistent use of weighting).

94. Nate Cohn, *How One 19-Year-Old Illinois Man Is Distorting National Polling Averages*, N.Y. Time, Oct. 12, 2016, <https://perma.cc/76R7-J363>.

designed to assess consumer confusion with Sun Time orange juice, potential respondents might be asked to name the brands of orange juice they have purchased recently or expect to purchase in the next six months. They should not be asked specifically if they recently have purchased, or expect to purchase, Sunshine orange juice, because this may affect their responses on the survey either by implying who is conducting the survey or by supplying them with a brand name that otherwise would not occur to them.

The content of a screening questionnaire (or “screener”) can also set the context for the questions that follow. In *Pfizer, Inc. v. Astra Pharmaceutical Products, Inc.*,⁹⁵ physicians were asked a screening question to determine whether they prescribed particular drugs. The survey question that followed the screener asked, “Thinking of the practice of cardiovascular medicine, what first comes to mind when you hear the letters XL?” The court found that the screener conditioned the physicians to respond with the name of a product (a drug) rather than a function (long acting).⁹⁶

The criteria for determining whether to include a potential respondent in the survey should be objective and clearly conveyed, preferably using written instructions addressed to those who administer the screening questions. These instructions and the completed screening questionnaire should be made available to the court and the opposing party along with the interview or survey form for each respondent. Computerized administration, described below, has allowed for this to be automated.

Survey Questions and Structure

Clarity, Precision, and Lack of Bias in the Framing of the Survey Questions

Although it seems obvious that questions on a survey should be clear and precise, phrasing questions to reach that goal is often difficult. Even questions that appear clear can convey unexpected meanings and ambiguities to potential respondents. For example, the question “What is the average number of days each week you have butter?” appears to be straightforward. Yet some respondents wondered whether margarine counted as butter, and when the question was revised to include the introductory phrase “not including margarine,” the reported frequency of butter use dropped dramatically.⁹⁷

95. 858 F. Supp. 1305, 1321 & n.13 (S.D.N.Y. 1994).

96. *Id.* at 1321.

97. Floyd J. Fowler, Jr., *How Unclear Terms Affect Survey Data*, 56 Pub. Op. Q. 218, 225–26 (1992), <https://doi.org/10.1086/269312>.

When unclear questions are included in a survey, they may threaten the validity of the survey by systematically distorting responses if respondents are misled in a particular direction, or by inflating random error if respondents guess because they do not understand the question.⁹⁸ If the crucial question is sufficiently ambiguous or unclear, it may be the basis for rejecting the survey. For example, a survey was designed to assess community sentiment that would warrant a change of venue in trying a case for damages sustained when a hotel skywalk collapsed.⁹⁹ The court found that the question “Based on what you have heard, read or seen, do you believe that in the current compensatory damage trials, the defendants, such as the contractors, designers, owners, and operators of the Hyatt Hotel, should be punished?” could neither be correctly understood nor easily answered.¹⁰⁰ The court noted that the phrase “compensatory damages,” although well-defined for attorneys, was unlikely to be meaningful for laypersons.¹⁰¹

Pilot work is a standard and valuable way to improve the quality of a survey.¹⁰² When there is any doubt whether a term or phrase will be clear to respondents, researchers should pretest to assess understanding, which may include focus groups or cognitive interviewing.¹⁰³

The value of pilot work is greatest when the issues and questions in the survey differ from the issues and questions addressed in previous surveys, or when the target population differs significantly from previously surveyed populations.¹⁰⁴

98. *See id.* at 219.

99. *Firestone v. Crown Ctr. Redevelopment Corp.*, 693 S.W.2d 99 (Mo. 1985) (en banc).

100. *See id.* at 102, 103.

101. *See id.* at 103. When there is any question about whether some respondents will understand a particular term or phrase, the term or phrase should be defined explicitly in the survey.

102. *See* Jon A. Krosnick & Stanley Presser, *Questions and Questionnaire Design*, in *Handbook of Survey Research*, *supra* note 1, at 294 (“No matter how closely a questionnaire follows recommendations based on best practices, it is likely to benefit from pretesting. . . .”). *See also* Jean M. Converse & Stanley Presser, *Survey Questions: Handcrafting the Standardized Questionnaire* 51 (1986); Fred W. Morgan, *Judicial Standards for Survey Research: An Update and Guidelines*, 54 J. Mktg. 59, 64 (1990), <https://doi.org/10.1177/002224299005400104>; OMB Standards and Guidelines for Statistical Surveys, Standard 1.4, Pretesting Survey Systems (2006), <https://perma.cc/D8XZ-4UX7> (specifying that to ensure that all components of a survey function as intended, pretests of survey components should be conducted unless those components have previously been successfully fielded); American Association for Public Opinion Research, *Best Practices* (2022), <https://perma.cc/Y3HU-XCHK> (“Before fielding a survey, it is important to pretest the questionnaire.”).

103. Cognitive interviewing includes a combination of think-aloud and verbal probing techniques. Gordon B. Willis et al., *Is the Bandwagon Headed to the Methodological Promised Land? Evaluating the Validity of Cognitive Interviewing Techniques*, in *Cognition and Survey Research* 136 (Monroe G. Sirken et al. eds., 1999). *See also* Gordon B. Willis, *Cognitive Interviewing in Survey Design: State of the Science and Future Directions*, in *The Palgrave Handbook of Survey Research* 103–07 (2018). *See also* Tourangeau et al., *supra* note 88, at 326–27.

104. Ivan R. Ross, *The Use of Pilot Tests and Pretests in Consumer Surveys*, in *Trademark and Deceptive Advertising Surveys* 13, 26 (Shari Diamond & Jerre Swann eds., 2d ed. 2022).

In many pretests or pilot tests,¹⁰⁵ the proposed survey is administered to a small sample (usually between twenty-five and seventy-five)¹⁰⁶ of the same type of respondents who would be eligible to participate in the full-scale survey. Some courts have explicitly recognized the value of pretests¹⁰⁷ and that lack of pretesting may suggest a weakness in the survey.¹⁰⁸

Litigants would be more likely to conduct pilot work and disclose it in expert reports if courts recognized that pilot work can maximize the likelihood that respondents understand the questions they are being asked. Moreover, the Federal Rules of Civil Procedure may require that a testifying expert disclose pilot work that serves as a basis for the expert's opinion. The situation is more complicated when a nontestifying expert conducts the pilot work and the testifying expert learns about the pilot testing only indirectly through the attorney's advice about the relevant issues in the case. Some commentators suggest that attorneys are obligated to disclose such pilot work.¹⁰⁹

What Respondents and Interviewers Knew About the Survey Purpose and Its Sponsorship

One way to protect the objectivity of survey administration is to avoid telling respondents or interviewers who is sponsoring the survey. Respondents who know the identity of the sponsor of the survey may adjust their responses, and interviewers who know the identity of the survey's sponsor may affect results inadvertently by communicating to respondents their expectations or what they believe are the preferred responses of the survey's sponsor. To ensure objectivity in the administration of the survey, it is standard interview practice in surveys conducted for litigation to do double-blind research whenever possible: Both the interviewer and the respondent are blind to the sponsor of the survey and its

105. The terms *pretest* and *pilot test* are sometimes used interchangeably to describe pilot work done in the planning stages of research. When they are distinguished, the difference is that a pretest tests the questionnaire, whereas a pilot test generally tests proposed collection procedures as well.

106. Converse & Presser, *supra* note 102, at 69. Converse and Presser suggest that a pretest with twenty-five respondents is appropriate when the survey uses professional interviewers.

107. See, e.g., *Zippo Mfg. Co. v. Rogers Imports, Inc.*, 216 F. Supp. 670 (S.D.N.Y. 1963); *Scott v. City of New York*, 591 F. Supp. 2d 554, 560 (S.D.N.Y. 2008) (“[T]he survey went through multiple pretests in order to insure its usefulness and statistical validity.”); *Estes Park Taffy Co. v. Original Taffy Shop, Inc.*, No. 15-cv-01697-CBS, 2017 U.S. Dist. LEXIS 88113, at *9 (D. Colo. June 8, 2017).

108. *GOLO, LLC. v. Goli Nutrition, Inc.*, No.20-667-RGA, 2020 U.S. Dist. LEXIS 158508, at *28 (D. Del. Sept. 1, 2020) (including lack of pretesting in a list of potential weaknesses in an expert's method).

109. See Yvonne C. Schroeder, *Pretesting Survey Questions: The Procedural and Ethical Ramifications*, 11 Am. J. Trial Advoc. 195, 197–201 (1987).

purpose. Thus, the survey instrument provides no explicit or implicit clues about the sponsorship of the survey (e.g., a sponsor's letterhead) or the expected responses (e.g., reversing the usual order of the yes and no response boxes on a key question, potentially increasing the likelihood that *no* will be checked).¹¹⁰

Nonetheless, in some situations (e.g., on some government surveys), disclosure of the survey's sponsor to respondents (and thus to interviewers) is required. Such surveys call for an evaluation of the likely biases introduced by interviewer or respondent awareness of the survey's sponsorship. In evaluating the consequences of sponsorship awareness, it is important to consider (1) whether the sponsor has views and expectations that are apparent and (2) whether awareness is confined to the interviewers or involves the respondents. For example, if a survey concerning attitudes toward gun control is sponsored by the National Rifle Association, it is clear that responses opposing gun control are likely to be preferred. In contrast, if the survey on gun control attitudes is sponsored by the Department of Justice, the identity of the sponsor may not suggest the kinds of responses the sponsor expects or would find acceptable.¹¹¹ When a survey involves interviewers who are well-trained, their awareness of sponsorship may be a less serious threat than respondents' awareness.¹¹²

In most situations, the survey expert can follow the traditional CASRO Code of Standards and Ethics¹¹³ and promise the respondent anonymity in the interest of obtaining the most accurate and candid responses. Moreover, in most situations, there is no need to disclose the identity of the survey sponsor. In some cases, however, respondents to a survey are involved in litigation (e.g., class-action wage and hour cases) and are likely to recognize that a survey asking them questions about relevant issues (such as unpaid overtime) is being sponsored by a party to the litigation. They may even be unwilling to participate in the survey unless they are told that the attorney representing them is sponsoring it. This creates a difficult problem for survey researchers. To get an acceptable response rate, it may be necessary to reveal who is sponsoring the survey. But this revelation may incentivize respondents to adjust their answers for potential financial gain, and promising anonymity may exacerbate this issue. If the survey is conducted by a friendly party, one approach that can be used to bolster accuracy is to tell respondents

110. See *Centaur Commc'ns, Ltd. v. A/S/M Commc'ns, Inc.*, 652 F. Supp. 1105, 1111 n.3 (S.D.N.Y.) (pointing out that reversing the usual order of response choices, "yes or no," to "no or yes" may confuse interviewers as well as introduce bias), *aff'd*, 830 F.2d 1217 (2d Cir. 1987).

111. See, e.g., Stanley Presser et al., *Survey Sponsorship, Response Rates, and Response Effects*, 73 Soc. Sci. Q. 699, 701 (1992) (different responses to a university-sponsored telephone survey and a newspaper-sponsored survey for questions concerning attitudes toward the mayoral primary, an issue on which the newspaper had taken a position).

112. See, e.g., Seymour Sudman et al., *Modest Expectations: The Effects of Interviewers' Prior Expectations on Responses*, 6 Soc. Methods & Rsch. 171, 181 (1977), <https://doi.org/10.1177/004912417700600203>.

113. CASRO, *supra* note 41, § I.A.

that their responses may not be anonymous and thus that exaggeration or other inaccuracies may be detected. It is unclear what the best practice is in this case, and the survey expert should be prepared to justify their choices.

Handling Respondents with No Opinions and Reducing Respondent Guessing

Some survey respondents may have no opinion about an issue under investigation, either because they have never thought about it before or because the question mistakenly assumes a familiarity with the issue. For example, some respondents in a consumer survey may not have noticed that the commercial they are being questioned about guaranteed the quality of the product being advertised, and thus they may have no opinion on the kind of guarantee it indicated. The following three alternative question structures will affect how those respondents answer and how their responses are counted.

First, the survey can ask all respondents to answer the question (e.g., “Did you understand the guarantee offered by Clover to be a 1-year guarantee, a 60-day guarantee, or a 30-day guarantee?”). Faced with a direct question, particularly one that provides response alternatives, the respondent obligingly may supply an answer even if (in this example) the respondent did not notice the guarantee. Such answers will reflect only what the respondent can glean from the question, or they may reflect pure guessing.

Second, the survey can use a quasi-filter question to reduce guessing by providing “don’t know” or “no opinion” options as part of the question (e.g., “Did you understand the guarantee offered by Clover to be for more than a year, a year, or less than a year, or don’t you have an opinion?”).¹¹⁴ By signaling to the respondent that it is acceptable not to have an opinion, the question reduces the demand for an answer and, as a result, the inclination to hazard a guess just to comply. Respondents are more likely to choose a “no opinion” option if it is mentioned explicitly by an interviewer than if it is merely accepted when the respondent spontaneously offers it as a response. Similarly, in an online survey, explicitly providing “don’t know” as a potential response rather than accepting it only if the respondent writes it in as an “other” response can increase the use of that option. The consequence of this change in format can be substantial. Studies indicate that, although the relative distribution of the respondents selecting the *listed* choices is unlikely to change dramatically, presentation of an explicit “don’t

114. Norbert Schwarz & Hans-Jürgen Hippler, *Response Alternatives: The Impact of Their Choice and Presentation Order*, in *Measurement Errors in Surveys* 41, 45–46 (Paul P. Biemer et al. eds., 1991); *Spangler Candy Co. v. Tootsie Roll Indus., LLC*, 372 F. Supp. 3d 588, 599 (N.D. Ohio 2019) (“Because [the expert] offered a ‘Don’t Know/No Opinion’ option for each closed-ended question, I conclude the questions were not unduly suggestive or guess-inducing.”).

know” or “no opinion” alternative commonly leads to a 20% to 25% increase in the proportion of respondents selecting that response.¹¹⁵

Finally, the survey can include full-filter questions—that is, questions that lay the groundwork for the substantive question by first asking respondents if they have an opinion about the issue or happened to notice the feature that the interviewer is preparing to ask about (e.g., “Based on the commercial you just saw, do you have an opinion about how long Clover stated or implied that its guarantee lasts?”).¹¹⁶ The survey then asks the substantive question only of those respondents who have indicated that they have an opinion on the issue.

Which of these three approaches is used and the way it is used can affect the rate of “no opinion” responses that the substantive question will evoke.¹¹⁷ Respondents are more likely to say that they do not have an opinion on an issue if a full filter is used than if a quasi-filter is used.¹¹⁸ However, in maximizing respondent expressions of “no opinion,” full filters may produce an underreporting of opinions. Some evidence indicates that full-filter questions discourage respondents who actually have opinions from offering them by conveying the implicit suggestion that respondents can avoid difficult follow-up questions by saying that they have no opinion.¹¹⁹

In general, then, a survey that uses full filters provides a conservative estimate of the number of respondents holding an opinion, while a survey that uses neither full filters nor quasi-filters may overestimate the number of respondents with opinions if some respondents offering opinions are guessing. The strategy of including a “no opinion” or “don’t know” response as a quasi-filter avoids both of these extremes, although some research suggests that even a quasi-filter may discourage a substantive answer from a respondent who would be able to provide one.¹²⁰

One solution that some survey researchers use is to provide respondents with a general instruction not to guess at the beginning of an interview, rather than supplying a “don’t know” or “no opinion” option as part of the options attached to each question.¹²¹ Another approach is to eliminate the “don’t know” option and to add follow-up questions that measure the strength of the respondent’s

115. Howard Schuman & Stanley Presser, *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording and Context* 113–46 (1996).

116. See, e.g., *Johnson & Johnson v. Merck Consumer Pharms. Co.*, 960 F.2d 294, 299 (2d Cir. 1992).

117. Considerable research has been conducted on the effects of filters. For a review, see George F. Bishop et al., *Effects of Filter Questions in Public Opinion Surveys*, 47 Pub. Op. Q. 528 (1983), <https://doi.org/10.1086/268810>.

118. Schwarz & Hippler, *supra* note 114, at 45–46.

119. *Id.* at 46.

120. Jon A. Krosnick et al., *The Impact of “No Opinion” Response Options on Data Quality: Non-Attitude Reduction or Invitation to Satisfice?*, 66 Pub. Op. Q. 371 (2002), <https://doi.org/10.1086/341394>.

121. *Anheuser-Busch, Inc. v. VIP Prods., LLC*, No. 4:08cv0358, 2008 U.S. Dist. LEXIS 82258, at *6 (E.D. Mo. Oct. 16, 2008).

opinion.¹²² More generally, survey experts can conduct pilot tests or cognitive interviews to assess whether many in the population in fact lack any opinion or the relevant knowledge to arrive at an opinion. If a substantial majority of people do have an opinion, it may be appropriate to exclude a “don’t know” option and thus avoid reducing data quality by encouraging less motivated respondents to use that response. In contrast, if many people do not have formed opinions or the knowledge required to arrive at them, the option should be included.¹²³ The survey expert should be prepared to explain why the “don’t know” option was included or excluded.

Use of Open-Ended and Closed-Ended Questions

The questions that make up a survey instrument may be open-ended, closed-ended, or a combination of both. Both are valid choices in appropriate situations. Open-ended questions require the respondent to formulate and express an answer in their own words (e.g., “What was the main point of the commercial?”). Closed-ended questions provide the respondent with an explicit set of responses from which to choose; the choices may be as simple as *yes* or *no* (e.g., “Is Colby College coeducational?”¹²⁴) or as complex as a range of alternatives (e.g., “The two pain relievers have (1) the same likelihood of causing gastric ulcers; (2) about the same likelihood of causing gastric ulcers; (3) a somewhat different likelihood of causing gastric ulcers; (4) a very different likelihood of causing gastric ulcers; or (5) none of the above.”¹²⁵). When a survey involves in-person interviews, the interviewer may show the respondent these choices on a showcard that lists them.

Open-ended and closed-ended questions may elicit very different responses.¹²⁶ Most responses are less likely to be volunteered by respondents who are asked an

122. Krosnick & Presser, *supra* note 102, at 285.

123. Dragana Bolcic-Jankovic et al., *Using “Don’t Know” Responses in a Survey of Oncologists Regarding Medicinal Cannabis*, 14 *Surv. Prac.* (2021), <https://doi.org/10.29115/SP-2020-0016>; see also Erika A. Waters et al., *Dismissing “Don’t Know” Responses to Perceived Risk Survey Items Threatens the Validity of Theoretical and Empirical Behavior-Change Research*, 17 *Persps. Psych. Sci.* 841 (2022), <https://doi.org/10.1177/17456916211017860>.

124. *President & Trs. of Colby College v. Colby College—N.H.*, 508 F.2d 804, 809 (1st Cir. 1975).

125. This question is based on one asked in *American Home Products Corp. v. Johnson & Johnson*, 654 F. Supp. 568, 581 (S.D.N.Y. 1987), that was found to be a leading question by the court, primarily because the choices suggested that the respondent had learned about aspirin’s and ibuprofen’s relative likelihoods of causing gastric ulcers. In contrast, in *McNeilab, Inc. v. American Home Products Corp.*, 501 F. Supp. 517, 525 (S.D.N.Y. 1980), the court accepted as nonleading the question, “Based only on what the commercial said, would Maximum Strength Anacin contain more pain reliever, the same amount of pain reliever, or less pain reliever than the brand you, yourself, currently use most often?”

126. Howard Schuman & Stanley Presser, *Question Wording as an Independent Variable in Survey Analysis*, 6 *Socio. Methods & Rsch.* 151 (1977), <https://doi.org/10.1177/004912417700600202>; Schuman & Presser, *supra* note 115, at 79–112; Converse & Presser, *supra* note 102, at 33.

open-ended question than they are to be chosen by respondents who are presented with a closed-ended question. The response alternatives in a closed-ended question may remind respondents of options that they would not otherwise consider or which simply do not come to mind as easily.¹²⁷

The advantage of open-ended questions is that they give the respondent fewer hints about expected or preferred answers. Response choices in a closed-ended question, in addition to reminding respondents of options that they might not otherwise consider, may direct the respondent away from or toward a particular response. For example, a commercial reported that in shampoo tests with more than 900 women, the sponsor's product received higher ratings than other brands.¹²⁸ According to a competitor, the commercial deceptively implied that each woman in the test rated more than one shampoo, when in fact each woman rated only one. To test consumer impressions, the survey showed the commercial and asked: "How many different brands mentioned in the commercial did each of the 900 women try?"¹²⁹ Respondents were given the choice of "one," "two," "three," "four," or "five or more." The fact that four of the five choices in the closed-ended question provided a response that was greater than one implied that the correct answer was probably more than one.¹³⁰

An open-ended question may also suggest that the answer is more than one, however. By asking "how many different brands," the question suggests (1) that the viewer should have received some message from the commercial about the number of brands each woman tried, and (2) that different brands were tried. Instead, a nonleading version of the question would have simply asked: "How many brands mentioned in the commercial did each of the 900 women try?"

127. For example, when respondents in one survey were asked, "What is the most important thing for children to learn to prepare them for life?", 62% picked "to think for themselves" from a list of five options, but only 5% spontaneously offered that answer when the question was open-ended. Schuman & Presser, *supra* note 115, at 104–07. An open-ended question presents the respondent with a free-recall task, whereas a closed-ended question is a recognition task. Recognition tasks in general reveal higher performance levels than recall tasks. Mary M. Smyth et al., *Cognition in Action* 25 (1987). In addition, there is evidence that respondents answering open-ended questions may be less likely to report some information that they would reveal in response to a closed-ended question when that information seems self-evident or irrelevant.

128. See *Vidal Sassoon, Inc. v. Bristol-Myers Co.*, 661 F.2d 272, 273 (2d Cir. 1981).

129. This was the wording of the closed-ended question in the survey discussed in *Vidal Sassoon*, 661 F.2d at 275–76.

130. Ninety-five percent of the respondents who answered the closed-ended question in the plaintiff's survey said that each woman had tried two or more brands. The open-ended question was never asked. *Vidal Sassoon*, 661 F.2d at 276. Norbert Schwarz, *Assessing Frequency Reports of Mundane Behaviors: Contributions of Cognitive Psychology to Questionnaire Construction*, in *Research Methods in Personality and Social Psychology* 98 (Clyde Hendrick & Margaret S. Clark eds., 1990), suggests that respondents often rely on the range of response alternatives as a frame of reference when they are asked for frequency judgments. See, e.g., Roger Tourangeau & Tom W. Smith, *Asking Sensitive Questions: The Impact of Data Collection Mode, Question Format, and Question Context*, 60 Pub. Op. Q. 275, 292 (1996), <https://doi.org/10.1086/297751>.

rather than “How many *different* brands” did each try? Similarly, an open-ended question that asks, “[W]hich company or store do you think puts out this shirt?” indicates to the respondent that the appropriate answer is the name of a company or store. The question would be leading if the respondent would have considered other possibilities (e.g., an individual or website if the question had not provided the frame of a company or store).¹³¹ Thus, the wording of a question, open-ended or closed-ended, can be leading, and the degree of suggestiveness of each question must be considered in evaluating the objectivity of a survey.

Closed-ended questions have some additional potential weaknesses that arise if the choices are not constructed properly. If the respondent is asked to choose one or more responses from among several choices, the response or responses chosen will be meaningful only if the list of choices is exhaustive—that is, if the choices cover all possible answers a respondent might give to the question. If the list of possible choices is incomplete, a respondent may be forced to choose one that does not express their opinion.¹³² The omission of a relevant choice option can only be partially attenuated by telling respondents explicitly that they are not limited to the choices presented, because most respondents nevertheless will select an answer from among the listed ones.¹³³

One form of closed-ended question format that can produce some distortion is the popular agree/disagree, true/false, or yes/no question. Although this format is appealing because it is easy to write and score these questions and their responses, the format is also seriously problematic. With its simplicity comes acquiescence: “[T]he tendency to endorse any assertion made in a question, regardless of its content,” is a systematic source of bias that has produced an inflation effect of 10% across a number of studies.¹³⁴ Only when control groups or control questions are added to the survey design, or when multiple items are counterbalanced, can this question format provide reasonable response estimates.¹³⁵

One challenge with open-ended responses is translating them into meaningful numeric metrics.¹³⁶ Imagine that a researcher asks respondents what they think about self-driving cars. If the researcher were interested in the prevalence and nature of safety concerns about self-driving cars, they would need a mechanism for translating the varied open-ended responses into a set of binary values: each possible safety concern was listed or not. This requires developing a list of different types of concerns such that each concern is unique (e.g., can malfunction, inability of driver to have control, etc.), and each has a text label. The list

131. *Smith v. Wal-Mart Stores, Inc.*, 537 F. Supp. 2d 1302, 1331–32 (N.D. Ga. 2008).

132. See, e.g., *Am. Home Prods. Corp. v. Johnson & Johnson*, 654 F. Supp. 568, 581 (S.D.N.Y. 1987).

133. See Howard Schuman, *Ordinary Questions, Survey Questions, and Policy Questions*, 50 Pub. Op. Q. 432, 435–36 (1986), <https://doi.org/10.1093/pubopq/50.3.432>.

134. Krosnick, *supra* note 86, at 552–53.

135. See section titled “Potential Order Effects” below.

136. Groves et al., *supra* note 77, at 332–44.

should also be exhaustive (cover all written answers) and include both a code for responses that do not refer to safety issues (e.g., reflects technological evolution) and also a code for nonresponses (e.g., left blank). In some cases, a respondent may give more than one answer, and that response should be coded in more than one category (e.g., “can malfunction and has no human control”). The researcher should be transparent on the construction of the coding system, which may be based on a priori categories or constructed based on the answers the respondents have given. Moreover, unless there are substantial data privacy concerns, the full open-ended answers should be made available to the other party for analysis.¹³⁷

The coding system should be clear enough that different coders would assign the same responses to the same categories in the vast majority of cases. In most academic research, this is accomplished by having multiple coders who are blind to the hypotheses and purpose of the project, giving them detailed instructions and training, and measuring inter coder reliability. Although this approach constitutes the best way to ensure reliable coding, because the raw data underlying the coding and the coding itself are made available to the court and the opposing party in litigation, the court and opposing party can directly examine and evaluate the appropriateness of the coding. It is therefore not uncommon for coding of simple measures in litigation to dispense with having multiple blind coders (though the resultant coding should be given close scrutiny).¹³⁸

Although many courts prefer open-ended questions on the ground that they are likely to be less leading, the value of any open-ended or closed-ended question depends on the information it conveys in the question and, in the case of closed-ended questions, in the choices provided. Open-ended questions are more appropriate when the survey is attempting to gauge what comes first to a respondent’s mind, but closed-ended questions are more suitable for assessing choices between well-identified options or obtaining ratings on a clear set of alternatives.

137. Mary B. Vardigan & Peter Granda, *Archiving, Documentation, and Dissemination*, in *Handbook of Survey Research* 707 (Peter V. Marsden & James D. Wright eds., 2d ed. 2010); see also section titled “Disclosure and Reporting” *infra*. Partial redaction of a response might be appropriate if the response was individually identifiable and the content sensitive. For example, a complaint about a named supervisor in an employment survey would fall in this category.

138. See, e.g., *Revlon Consumer Prods. Corp. v. Jennifer Leather Broadway, Inc.*, 858 F. Supp. 1268, 1276 (S.D.N.Y. 1994) (inconsistent scoring and subjective coding led court to find survey so unreliable that it was entitled to no weight), *aff’d*, 57 F.3d 1062 (2d Cir. 1995); *Rock v. Zimmerman*, 959 F.2d 1237, 1253 n.9 (3d Cir. 1992) (court found that responses on a change-of-venue survey incorrectly categorized respondents who believed the defendant was insane as believing he was guilty); *Coca-Cola Co. v. Tropicana Prods., Inc.*, 538 F. Supp. 1091, 1094–96 (S.D.N.Y.) (plaintiff’s expert stated that respondents’ answers to the open-ended questions revealed that 43% of respondents thought Tropicana was portrayed as fresh-squeezed; the court’s own tabulation found no more than 15% believed this was true), *rev’d on other grounds*, 690 F.2d 312 (2d Cir. 1982); see also *Cumberland Packing Corp. v. Monsanto Co.*, 140 F. Supp. 2d 241 (E.D.N.Y. 2001) (court examined verbatim responses that respondents gave to arrive at a confusion level substantially lower than the level reported by the survey expert).

Use of Probes to Clarify Ambiguous or Incomplete Answers

When questions allow respondents to express their opinions in their own words, some of the respondents may give ambiguous or incomplete answers. If the survey is administered online, some probes (e.g., “Why did you choose that answer?”) can be programmed into the survey instrument and automatically administered as part of the survey. If interviewers are asking the questions, they may be instructed to probe to obtain a more complete response or clarify the meaning of the ambiguous response. Interviewers should be instructed what clarification, if any, they can provide. The record should reflect both what the respondent said initially, what the interviewer said in the attempt to get or provide clarification, and how the respondent answered the probe; this information will allow the court and the opposing party to evaluate whether the probe affected the views expressed by the respondent.

If the survey involves interviewers who are permitted to administer follow-up questions, they must be given explicit instructions on when they should probe and what they should say in probing.¹³⁹ Standard probes used to draw out all that the respondent has to say (e.g., “Any further thoughts?”; “Anything else?”; “Can you explain that a little more?”; or “Could you say that another way?”) are relatively noncontroversial and can be programmed into an online survey instrument. But probing should be limited. Persistent continued requests for further responses to the same or nearly identical questions may convey the idea to the respondent that they have not yet produced the “right” answer, particularly if the probes are administered by a live interviewer.¹⁴⁰

Potential Order Effects

The order in which questions are asked on a survey and the order in which response alternatives are provided in a closed-ended question can influence the answers.¹⁴¹ For example, although asking a general question before a more specific

139. Floyd J. Fowler, Jr. & Thomas W. Mangione, *Standardized Survey Interviewing: Minimizing Interviewer-Related Error* 41–42 (1990), <https://doi.org/10.4135/9781412985925>.

140. *See, e.g., Johnson & Johnson-Merck Consumer Pharms. Co. v. Rhone-Poulenc Rorer Pharms., Inc.*, 19 F.3d 125, 135 (3d Cir. 1994); *Am. Home Prods. Corp. v. Procter & Gamble Co.*, 871 F. Supp. 739, 748 (D.N.J. 1994).

141. *See* Schuman & Presser, *supra* note 115, at 23, 56–74; Krosnick & Presser, *supra* note 102, at 278–81. In *R.J. Reynolds Tobacco Co. v. Loew's Theatres, Inc.*, 511 F. Supp. 867, 875 (S.D.N.Y. 1980), the court recognized the biased structure of a survey that disclosed the tar content of the cigarettes being compared before questioning respondents about their cigarette preferences. Not surprisingly, respondents expressed a preference for the lower tar product. *See also* *E. & J. Gallo Winery v. Pasatiempos Gallo, S.A.*, 905 F. Supp. 1403, 1409–10 (E.D. Cal. 1994) (court recognized

question on the same topic is unlikely to affect the response to the specific question, reversing the order of the questions may influence responses to the general question. As a rule, then, surveys are less likely to be subject to order effects if the questions move from the general (e.g., “What do you recall being discussed in the advertisement?”) to the specific (e.g., “Based on your reading of the advertisement, what companies do you think the ad is referring to when it talks about rental trucks that average five miles per gallon?”).¹⁴²

The mode of questioning can influence the form that an order effect takes. When respondents are shown response alternatives visually, as in mail surveys and self-administered online surveys or in face-to-face interviews when respondents are shown a card containing response alternatives, they are more likely to select the first choice offered (a primacy effect).¹⁴³ In contrast, when response alternatives are presented orally, as in telephone surveys, respondents are more likely to choose the last choice offered (a recency effect).¹⁴⁴ Although these effects are typically small, no general formula is available that can adjust values to correct for order effects, because the size and even the direction of the order effects may depend on the nature of the question being asked and the choices being offered. To control for order effects, the order of the response choices in a survey should be rotated, randomized, or counterbalanced,¹⁴⁵ so that no response alternative will have an inflated chance of being selected because of its position.

Appropriate Inclusion of Control Groups, Control Questions, or Other Comparisons

Many surveys are designed not simply to describe attitudes, beliefs, or reported behaviors, but to determine the source of those attitudes, beliefs, or behaviors. That is, the purpose of the survey is to test a causal proposition. For example, how does a trademark or the content of a commercial affect respondents’ perceptions or understanding of a product or commercial? Thus, the question is not merely whether consumers hold inaccurate beliefs about Product A, but whether exposure to the commercial misleads the consumer into thinking that Product A

that earlier questions referring to playing cards, board or table games, or party supplies, such as confetti, increased the likelihood that respondents would include these items in answers to the questions that followed).

142. This question was accepted by the court in *U-Haul International, Inc. v. Jartran, Inc.*, 522 F. Supp. 1238, 1249 (D. Ariz. 1981), *aff’d*, 681 F.2d 1159 (9th Cir. 1982).

143. Krosnick & Presser, *supra* note 102, at 280.

144. *Id.*

145. See, e.g., *Winning Ways, Inc. v. Holloway Sportswear, Inc.*, 913 F. Supp. 1454, 1465–67 (D. Kan. 1996) (failure to rotate the order in which the jackets were shown to the consumers led to reduced weight for the survey); *Procter & Gamble Pharms., Inc. v. Hoffmann-La Roche, Inc.*, No. 06 Civ. 0034 (PAC), 2006 U.S. Dist. LEXIS 64363 (S.D.N.Y. Sept. 6, 2006).

is a superior pain reliever. Yet if consumers already believe, before viewing the commercial, that Product A is a superior pain reliever, a survey that simply records consumers' impressions after they view the commercial may reflect those preexisting beliefs rather than impressions produced by the commercial.

Some surveys attempt to reduce the impact of preexisting impressions on respondents' answers by instructing respondents to focus solely on the stimulus as a basis for their answers. Thus, the survey includes a preface (e.g., "based on the commercial you just saw") or directs the respondent's attention to the mark at issue (e.g., "these stripes on the package"). Such efforts are likely to be only partially successful. It is often difficult for respondents to identify accurately the source of their impressions.¹⁴⁶ The more routine the idea being examined in the survey, the more likely it is that the respondent's answer is influenced by (1) preexisting impressions; (2) general expectations about what commercials typically say (e.g., the product being advertised is better than its competitors); or (3) guessing, rather than by the actual content of the commercial message or the trademark being evaluated. A similar limitation occurs when respondents are asked to explain why they chose a particular response: they can justify their choice, but may not be able to report accurately what actually led them to make that choice.

With respondents randomly assigned to one or more appropriate comparison conditions, the survey expert can draw clear causal inferences about the influence of the stimulus.¹⁴⁷ In the simplest version of such a survey-experiment (as discussed above), respondents are assigned randomly to one of two conditions.¹⁴⁸ For example, respondents assigned to the experimental condition view an allegedly deceptive commercial, and respondents assigned to the control condition view the same commercial with the allegedly deceptive material removed or an alternative commercial that does not contain the allegedly deceptive material.¹⁴⁹

146. See Richard E. Nisbett & Timothy D. Wilson, *Telling More Than We Can Know: Verbal Reports on Mental Processes*, 84 *Psych. Rev.* 231 (1977), <https://doi.org/10.1037/0033-295X.84.3.231>.

147. See Shari S. Diamond, *Using Psychology to Control Law: From Deceptive Advertising to Criminal Sentencing*, 13 *L. & Hum. Behavior* 239, 244–46 (1989), <https://doi.org/10.1007/BF01067028>; Jacob Jacoby & Constance Small, *Applied Marketing: The FDA Approach to Defining Misleading Advertising*, 39 *J. Mktg.* 65, 68 (1975), <https://doi.org/10.1177/002224297503900413>. See also R. Charles Henn, *Why Ask Why? A Critical Assessment of an Historical Survey Artifact*, 113 *Trademark Rep.* 772, 773–76 (2023).

148. Random assignment should not be confused with random selection. When respondents are assigned randomly to different treatment groups (e.g., respondents in each group watch a different commercial), the procedure ensures that within the limits of sampling error the two groups of respondents will be equivalent, on average, except for the different treatments they receive. Respondents selected for a mall intercept study, and not from a probability sample, may be assigned randomly to different treatment groups. Random selection, in contrast, describes the method of selecting a sample of respondents in a probability sample. See section titled "The Sample as a Reflection of the Relevant Characteristics of the Population" above.

149. This alternative commercial could be a "tombstone" advertisement that includes only the name of the product or a more elaborate commercial that does not include the claim at issue.

Respondents in both the experimental and control groups answer the same set of questions about the allegedly deceptive message. The effect of the commercial's allegedly deceptive message is evaluated by comparing the responses made by the experimental group members with those of the control group members. If 40% of the respondents in the experimental group responded indicating a belief in the deceptive claim, whereas only 8% of the respondents in the control group gave that response, the difference between 40% and 8% (within the limits of sampling error¹⁵⁰) can be attributed to the allegedly deceptive commercial.

A survey-experimental design with an appropriate control group can account for more than preexisting beliefs.¹⁵¹ Other sources of systematic and random error can influence responses to survey questions. Thus, if a respondent is asked "Have you heard of a security company named Titan Alarm?" some individuals will say they have, even if they have not. A control group can help address this type of error as well; this and other background noise should have produced similar response levels in the experimental and control groups, so comparing average scores in those groups should control for those sources of error. In addition, a leading question may cause participants to be more likely to endorse or reject a particular statement. But if respondents who viewed the allegedly deceptive commercial respond differently than respondents who viewed the control commercial, the difference cannot be merely the result of a leading question, because both groups answered the same question. The ability to evaluate the effect of the wording of a particular question makes the control group design particularly useful in assessing responses to closed-ended questions,¹⁵² which may encourage guessing or particular responses. Thus, the focus is not on the absolute response level in the experimental group, but on the difference between the responses from the experimental group and those from the control group.¹⁵³

In designing a survey-experiment, the expert should select a stimulus for the control group that shares as many characteristics with the experimental stimulus as possible, with the key exception of the characteristic whose influence is being assessed.¹⁵⁴ Although a survey with an imperfect control group may provide

150. For a discussion of sampling error, see the glossary to the current chapter and David H. Kaye and Hal S. Stern, *Reference Guide on Statistics and Research Methods*, in this manual.

151. For a more extensive discussion of controls, see Shari Seidman Diamond, *Control Foundations: Rationale and Approaches*, in *Trademark and Deceptive Advertising Surveys* 239 (Shari Diamond & Jerre Swann eds., 2d ed. 2022).

152. The Federal Trade Commission has long recognized the need for some kind of control for closed-ended questions, although it has not specified the type of control that is necessary. See *In re Stouffer Foods Corp.*, 118 F.T.C. 746, 808–09 (Sept. 26, 1994).

153. See, e.g., *CytoSport, Inc. v. Vital Pharms., Inc.*, 617 F. Supp. 2d 1051, 1075–76 (E.D. Cal. 2009) (net confusion level of 25.4% obtained by subtracting 26.5% in the control group from 51.9% in the test group).

154. See, e.g., *Skechers USA, Inc. v. Vans, Inc.*, No. CV-07-01703 DSF (PLAx), 2007 WL 4181677, at *8–9 (C.D. Cal. Nov. 20, 2007) (in trade dress infringement case, control stimulus should have retained design elements not at issue); *Procter & Gamble Pharms., Inc. v. Hoffman-LaRoche*,

better information than a survey with no control group at all, the choice of an appropriate control group requires care and should influence the admissibility of the survey and the weight that the survey receives. For example, a control stimulus should not be less attractive than the experimental stimulus if the survey is designed to measure how familiar the experimental stimulus is to respondents, because attractiveness may affect perceived familiarity.¹⁵⁵ Nor should the control stimulus share with the experimental stimulus the feature whose impact is being assessed. If the control stimulus in a case of alleged trademark infringement is itself a likely source of consumer confusion, for example, reactions to the experimental and control stimuli may not differ because both cause respondents to express a similar level of confusion.¹⁵⁶ In an extreme case, an inappropriate control may do nothing more than control for the effect of the nature or wording of the survey questions (e.g., acquiescence).¹⁵⁷ That generally will not be enough to rule out other explanations for different or similar responses to the experimental and control stimuli. Finally, it may sometimes be appropriate to have more than one control group to assess precisely what is causing the response to the experimental stimulus (e.g., in the case of an allegedly deceptive ad, whether it is a misleading graph or a misleading claim by the announcer, or in the case of allegedly infringing trade dress, whether it is the style of the font used or the coloring of the packaging).¹⁵⁸

Courts have increasingly come to recognize the central role that control groups play in evaluating causal claims.¹⁵⁹ Litigants have taken the cue, and most

Inc., No. 06 Civ 0034 (PAC), 2006 U.S. Dist. LEXIS 64363, at *87–88 (S.D.N.Y. Sept. 6, 2006) (in false advertising action, disclaimer was inadequate substitute for appropriate control group).

155. See, e.g., *Indianapolis Colts, Inc. v. Metro. Balt. Football Club L.P.*, 34 F.3d 410, 415–16 (7th Cir. 1994) (court recognized that the name “Baltimore Horses” was less attractive for a sports team than the name “Baltimore Colts”); see also *Reed-Union Corp. v. Turtle Wax, Inc.*, 77 F.3d 909, 912 (7th Cir. 1996) (court noted that one expert’s choice of a control brand with a well-known corporate source was less appropriate than the opposing expert’s choice of a control brand whose name did not indicate a specific corporate source).

156. See, e.g., *W. Publ’g Co. v. Publ’ns Int’l, Ltd.*, No. 94–C–6803, 1995 U.S. Dist. LEXIS 5917, at *45 (N.D. Ill. May 2, 1995) (court noted that the control product was “arguably more infringing than” the defendant’s product) (emphasis omitted). See also *Classic Foods Int’l Corp. v. Kettle Foods, Inc.*, No. SACV 04–725 CJC (Ex), 2006 U.S. Dist. LEXIS 97200 (C.D. Cal. Mar. 2, 2006); *McNeil-PPC, Inc. v. Merisant Co.*, No. 04–1090 (JAG), 2004 U.S. Dist. LEXIS 27733 (D.P.R. July 29, 2004).

157. See *supra* text accompanying note 134.

158. See, e.g., *Masterfoods USA v. Arcor USA, Inc.*, 230 F. Supp. 2d 302 (W.D.N.Y. 2002).

159. See, e.g., *Colangelo v. Champion Petfoods USA, Inc.*, No. 6:18–CV–1228, 2022 U.S. Dist. LEXIS 60489, at *32 (N.D.N.Y. Mar. 31, 2022) (“[W]ithout a control group it is impossible to establish cause and effect.”); *Longoria v. Million Dollar Corp.*, No. 18–CV–02266–PAB–NYW, 2021 U.S. Dist. LEXIS 38478, at *27 (D. Colo. Mar. 2, 2021) (survey excluded because “a causal study . . . requires a control group”); *SmithKline Beecham Consumer Healthcare, L.P. v. Johnson & Johnson-Merck*, No. 01 Civ. 2775 (DAB), 2001 U.S. Dist. LEXIS 7061, at *37–39 (S.D.N.Y. June 1, 2001) (survey to assess implied falsity of a commercial not probative in the absence of a control group);

experts recognize the need to submit a survey with a control group (i.e., a survey-experiment) if they wish to make a causal claim and rule out other explanations for the responses to the survey questions.¹⁶⁰

A less common control methodology is a control question. Rather than administering a control stimulus to a separate group of respondents, the survey asks all respondents one or more control questions along with the question about the product or service at issue. This technique is used to evaluate whether a brand name is generic. The genericness survey presents survey respondents with a series of product or service names and asks them to indicate in each instance whether they believe the name is a brand name or a common name. By showing that 68% of respondents considered Teflon a brand name (a proportion similar to the 75% of respondents who recognized the acknowledged trademark Jell-O as a brand name, and markedly different from the 13% who thought aspirin was a brand name), the makers of Teflon demonstrated that their respondents understood the difference between brands and product categories (so-called common names) and that the Teflon mark at issue was perceived as a brand name; they retained their trademark.¹⁶¹ It is crucial to control for order effects in such designs.

The issue of appropriate comparisons is especially salient in the context of conjoint analysis, which has been used in some patent and false-advertising cases. In cases of patent infringement, the plaintiff must submit an estimate of the damages produced by the defendant's infringement. When the infringement involves a multicomponent product, the plaintiff must apportion damages to determine the value of the patented component at issue.¹⁶² Similarly, in class actions involving false advertising, damages depend on the value to consumers that can be attributed to the deceptive portion of the advertisement.¹⁶³ To estimate these damages, survey experts have increasingly turned to conjoint analysis. When done

Am. Home Prods. Corp. v. Procter & Gamble Co., 871 F. Supp. 739, 749 (D.N.J. 1994) (discounting survey results based on failure to control for participants' preconceived notions); *ConAgra, Inc. v. Geo. A. Hormel & Co.*, 784 F. Supp. 700, 728 (D. Neb. 1992) ("Since no control was used, the . . . study, standing alone, must be significantly discounted."), *aff'd*, 990 F.2d 368 (8th Cir. 1993).

160. William, R. Shadish et al., *Experimental and Quasi-Experimental Designs for Generalized Causal Inference* (2002); James N. Druckman, *Experimental Thinking: A Primer on Social Science Experiments* (2022).

161. *E.I. DuPont de Nemours & Co. v. Yoshida Int'l, Inc.*, 393 F. Supp. 502, 526–27 & n.54 (E.D.N.Y. 1975); *see also* *Donchez v. Coors Brewing Co.*, 392 F.3d 1211, 1218 (10th Cir. 2004) (respondents evaluated eight brand and generic names in addition to the disputed name); *Reinalt-Thomas Corp. v. Mavis Tire Supply, LLC*, 391 F. Supp. 3d 1261, 1271–73 (N.D. Ga. 2019). A similar approach is used in assessing secondary meaning. *See, e.g., T-Mobile US, Inc. v. AIO Wireless LLC*, 991 F. Supp. 2d 888 (S.D. Tex. 2014).

162. For example, in *Apple, Inc. v. Samsung Electronics Co., Ltd.*, No. 11-CV-01846-LHK, 2014 U.S. Dist. LEXIS 29721 (N.D. Cal. Mar. 6, 2014), Apple used a conjoint survey to isolate the costs associated with Samsung's alleged patent infringement regarding iPhone and iPad features.

163. *E.g., Price v. L'Oréal U.S., Inc.*, No. 17-Civ-614 (LGS), 2020 Dist. LEXIS 153255 (S.D.N.Y. Aug. 24, 2020).

well, conjoint analysis can provide useful information, but there are many decisions made in designing a conjoint analysis that can undermine reliability and bias the resulting estimates.¹⁶⁴

In contrast to a survey that asks respondents directly how much they would be willing to pay (WTP) for a given feature or attribute of a product, a task that may be difficult for consumers, a conjoint survey-experiment typically presents respondents with choices between products with different randomly assigned profiles consisting of combinations of features.¹⁶⁵ Suppose the survey is designed to determine the value of a patented component of a washing machine that shuts off the machine automatically in response to overflow (automatic shutoff). The respondent will be presented with a series of choices between washing machines that may, for example, vary on brand (LG, General Electric, Whirlpool), number of settings (eight, ten, twelve), color (white, beige, stainless steel), capacity (3, 4, 4.5 cu. ft.), and price (\$400, \$598, \$648). Each respondent makes their choice in response to several sets (profiles) of products where the attributes (e.g., brand, number of settings) have randomly assigned levels (e.g., LG, General Electric, Whirlpool; eight, ten, twelve) for each profile. A regression is then used to compute part worths, the value that each feature adds to the total product. As in any survey, it is crucial to identify the appropriate universe of individuals likely to purchase the product. For instance, it would be inappropriate to conduct the washing machine conjoint survey on college students who can be assumed to have no intention of purchasing a washing machine in the near future.

Although conjoint analysis offers a potentially useful method for assessing WTP for a single feature of a multifeature product as a result of its experimental nature, the design of the profiles can lead to misleading results. The first issue is what features to include in the product profiles. If important features of the product are omitted (e.g., in the washing machine example, whether the machine is front-loading or top-loading) and the profiles include predominantly less important features e.g., door shape as round versus square, maximum spin speed), the estimated values for the component of interest are likely to be inflated.¹⁶⁶ Thus, an important preliminary step by the expert designing a conjoint analysis is to assess the importance of the various features of the product.¹⁶⁷

164. Bernard Chao & Sydney Donovan, *Does Conjoint Analysis Reliably Value Patents?*, 58 Am. Bus. L.J. 225 (2021), <https://doi.org/10.1111/ablj.12182>. See also Suneal Bedi & David Reibstein, *Damaged Damages: Errors in Patent and False Advertising Litigation*, 73 Ala. L. Rev. 385 (2021); David Franklyn & Adam Kuhn, *The Problem of Mop Heads in the Era of Apps: Toward More Rigorous Standards of Value Apportionment in Contemporary Patent Law*, 98 J. Pat. & Trademark Off. Soc'y 182 (2016).

165. Some conjoint analyses ask respondents for rankings or ratings, but asking them to make choices is the most generally accepted approach because it is more reflective of what they would do in the marketplace. Bedi & Reibstein, *supra* note 164, at 399 nn.76 & 77.

166. Bedi & Reibstein, *supra* note 164.

167. In *Apple, Inc. v. Samsung Electronics Co.*, No. 11-CV-01846-LHK, 2014 U.S. Dist. LEXIS 29721 (N.D. Cal. Mar. 6, 2014), the conjoint survey used by Apple included seven attributes;

An additional challenge in designing a conjoint survey is to ensure that the features are described clearly and accurately, while simultaneously not drawing substantially more attention to features than consumers might naturally devote to them in the real world. This allows respondents to understand what each feature is and ensures that their understanding matches the meanings and definitions used in the case.¹⁶⁸ If a feature is unclear (e.g., electronic spinner), the respondent cannot make an informed decision that takes that feature into account. It also is essential, as with any survey, that the population be carefully chosen.¹⁶⁹

Further, if the conjoint analysis is being used to identify prices, there is some controversy on how to capture supply-side aspects of the pricing that reflect the market. Given that market conditions may have changed due to the infringement, this can present a challenge.¹⁷⁰

Samsung critiqued it for excluding other important attributes such as brand name and battery life. In *MacDougall v. American Honda Motor Co.*, No. 20–56060, 2021 U.S. App. LEXIS 37780 (9th Cir. Dec. 21, 2021), the court of appeals reversed after the district court had rejected the use of a conjoint survey for “the reduction of the amount of vehicle attributes in the final survey from thirty-three to four The more limited a consumer’s choice of vehicle features, the more artificially inflated the importance of the remaining features. . . .” *MacDougall v. Am. Honda Motor Co.*, No. SACV 17–1079 JGB, 2020 U.S. Dist. LEXIS 166786, at *22 (C.D. Cal., Sept. 11, 2020). Including thirty-three attributes may be too many, but at issue here was the lack of justification for the four chosen (it was unclear how the pretest led to choosing those four). On appeal, the court found that the attributes chosen go to weight and not admissibility; however, the lack of a reasonable basis for attribute choice can seriously undermine the value of the analysis. There often is an inevitable tradeoff between overwhelming respondents with too many attributes and capturing those that are most important. The expert should explicitly provide reasons for the specific inclusion and exclusion of attributes.

168. For instance, in *Price*, the court rejected an analysis that used the term “Kerantindose Pro-Keratin+Silk” to isolate the impact of “Kerantindose” and “Pro-Keratin” since it includes “+Silk.” 2020 Dist. LEXIS 153255. In *Allegra v. Luxottica Retail North America*, 341 F.R.D. 373 (E.D.N.Y. Dec. 13, 2021), the court rejected the use of a conjoint survey that it said failed to properly describe the allegedly fraudulent omission in the advertising.

169. In *Cardenas v. Toyota Motor Corp.*, 418 F. Supp. 3d 1090 (S.D. Fla. 2019), the case involved a defect in a nonhybrid Toyota. The defendants argued that hybrid purchasers should not have been included. The crucial question was whether purchasers of hybrid cars would differ in their assessments from nonhybrid purchasers, which the court determined was a question of weight rather than admissibility. The analysis of subgroup effects is tricky because respondent subgroups may differ in the evaluations of comparison points and thus analyses needed to account for those distinctions; see Thomas Leeper et al., *Measuring Subgroup Preferences in Conjoint Experiments*, 28 Pol. Analysis, 207–21 (2020), <https://doi.org/10.1017/pan.2019.30>.

170. J. Gregory Sidak & Jeremy O. Skog, *Using Conjoint Analysis to Apportion Patent Damages*, 25 Fed. Cir. Bar J. 581 (2016). More generally, how to estimate supply side considerations is a matter of some debate. In *Cardenas v. Toyota Motor Corp.*, the court acknowledged differing opinions on whether the appropriate pricing information should emulate actual market prices and sales during the class period or the probable prices and sales in the situation where the relevant attribute took on the value that was under contention (the relevant attribute in this case was an odor emitted from a car’s HVAC system). The court ruled that the approach to incorporating supply-side information does not go to admissibility, but could affect weight. See also *Colangelo v. Champion*

Some aspects of conjoint analysis are controversial. These include whether it is always necessary to give respondents the choice of declining to select any of the products offered in the profiles presented (i.e., “none of the above”). If the relevant population includes people who may prefer not to make the purchase, inclusion is warranted, but note there is a danger that respondents may choose the “no purchase” option to avoid assessing the products.¹⁷¹ Similarly, there is some dispute about how many features of a product can be listed (e.g., can respondents process and respond to more than seven features?).

As with any survey, the expert should be prepared to explain the choices made in constructing the survey design. In light of the complexity of conjoint survey experimental designs and the controversial nature of some of the decisions the expert needs to make, a detailed explanation of those decisions is warranted to guide the court in evaluating whether the survey is admissible (or what weight to give it).

Benefits and Limitations Associated with the Mode of Data Collection Used in the Survey

Four primary methods have traditionally been used to collect survey data: (1) in-person interviews, (2) telephone interviews, (3) mail questionnaires, and (4) internet surveys. The choice of any data collection method for a survey should be justified by its strengths and weaknesses.

Common across in-person, telephone, and internet surveys is the use of computer-assisted techniques.¹⁷² The interviewer conducting a computer-assisted interview (CAI), whether by telephone (CATI) or face-to-face (computer-assisted personal interviewing, or CAPI), follows the computer-generated script for the interview and enters the respondent’s answers as the interview proceeds. A primary advantage of CATI and other CAI procedures is that skip patterns can be built into the program. If, for example, the respondent answers “yes” when asked whether she has ever been the victim of a burglary, the computer will generate further questions about the burglary; if she answers

Petfoods USA, Inc., No. 6:18-CV-122 (LEK/ML), 2022 U.S. Dist. LEXIS 60489, at *25 (N.D.N.Y. Mar. 31, 2022) (“[T]here is a legitimate difference of opinions, both among judges and experts, about the significance of supply side information in calculating loss of value.”) (citing *In re Fisher-Price Rock ‘n’ Play Sleeper Mktg.*, 567 F. Supp. 3d 406, 415 (W.D.N.Y. 2021)).

171. Daniel McFadden, *Stated Preference Methods and Their Applicability to Environmental Use and Non-use Valuations*, in *Contingent Valuation of Environmental Goods: A Comprehensive Critique*, 153–87 (Daniel McFadden & Kenneth Train eds., 2017). See also Moshe Ben-Akiva et al., *Foundations of Stated Preference Elicitation: Consumer Behavior and Choice-Based Conjoint Analysis*, 10 *Founds. & Trends in Econometrics* 1 (2019), <http://dx.doi.org/10.1561/08000000036>.

172. Wright & Marsden, *supra* note 1, at 13–14.

no, the program will automatically skip the follow-up burglary questions. Interviewer errors in following the skip patterns are therefore avoided, making CAI procedures particularly valuable when the survey involves complex branching and skip patterns.¹⁷³ CAI procedures also can be used to control for order effects by having the program rotate the order in which the questions or choices are presented and facilitate the implementation of complex experiments with many conditions by randomly generating many potential factors at once.¹⁷⁴

CAI procedures also include audio computer-assisted self-interviewing (ACASI) in which the respondent listens to recorded questions over the telephone or reads questions from a computer screen while listening to recorded versions of them through headphones. The respondent then answers verbally or directly on a keypad. ACASI procedures are particularly useful for collecting sensitive information (e.g., illegal drug use and other HIV risk behavior).¹⁷⁵ This is also useful in face-to-face interviews where the respondent can avoid having to reveal their (potentially sensitive) answer to the interviewer since they enter it themselves.

All CAI procedures require additional planning to take advantage of the potential for improvements in data quality. When a CAI protocol is used in a survey presented in litigation, the party offering the survey should supply for inspection the computer program that was used to generate the interviews. Moreover, CAI procedures do not eliminate the need for close monitoring of interviews to ensure that interviewers are accurately reading the questions in the interview protocol and accurately entering the respondent's answers (or that the respondent is paying attention and using the program correctly in entering their own answers).

In-Person Interviews

Although costly, in-person interviews generally are the preferred method of data collection, especially when visual materials must be shown to the respondent under controlled conditions. When the questions are complex and the interviewers are skilled, in-person interviewing provides the maximum opportunity to clarify or probe. Unlike a mail survey, in-person, telephone, and web-based surveys have the capability to implement complex skip sequences (in which the respondent's answer determines which question will be asked next) and the power to control the order in which the respondent answers the questions. In-person

173. Willem E. Saris, *Computer-Assisted Interviewing* 20, 27 (1991).

174. See, e.g., *Intel Corp. v. Advanced Micro Devices, Inc.*, 756 F. Supp. 1292, 1296–97 (N.D. Cal. 1991) (survey designed to test whether the term *386* as applied to a microprocessor was generic used a CATI protocol that tested reactions to five terms presented in rotated order).

175. See, e.g., P.C. Cooley et al., *Automating Telephone Surveys: Using T-ACASI to Obtain Data on Sensitive Topics*, 16 *Computs. in Hum. Behav.* 1 (2000), <https://perma.cc/L94R-AXKY>.

interviewers also can directly verify who is completing the survey, a check that is unavailable in mail and web-based surveys. As described below in the “Selection and Training of the Interviewers” section, appropriate interviewer training, as well as monitoring of the implementation of interviewing, is necessary if these potential benefits are to be realized. Objections to the use of in-person interviews arise primarily from their high cost or, on occasion, from evidence of inept or biased interviewers. The latter concern is somewhat mitigated by computer assistance, described above.

Telephone Interviews

Telephone surveys offer a comparatively fast and lower-cost alternative to in-person surveys and can be particularly useful when the population is large and geographically dispersed. Telephone interviews (unless supplemented with mailed materials) should be used only when it is unnecessary to show the respondent any visual materials. Thus, an attorney may present the results of a telephone survey of jury-eligible citizens in a motion for a change of venue in order to provide evidence that community prejudice raises a reasonable suspicion of potential jury bias.¹⁷⁶ Similarly, potential confusion between a restaurant called McBagel’s and the McDonald’s fast-food chain was established in a telephone survey. Over objections from defendant McBagel’s that the survey did not show respondents the defendant’s print advertisements, the court found likelihood of confusion based on the survey, noting that “by soliciting audio responses [the telephone survey] was closely related to the radio advertising involved in the case.”¹⁷⁷ In contrast, when words are not sufficient because, for example, the survey is assessing reactions to the trade dress or packaging of a product that is alleged to promote confusion, a telephone survey alone does not offer a suitable vehicle for questioning respondents.¹⁷⁸

176. See, e.g., *State v. Baumruk*, 85 S.W.3d 644 (Mo. 2002) (overturning the trial court’s decision to ignore a survey that found about 70% of county residents remembered the shooting that led to the trial, and that of those who had heard about the shooting, 98% believed that the defendant was either definitely guilty or probably guilty); *State v. Erickstad*, 620 N.W.2d 136, 140 (N.D. 2000) (denying change-of-venue motion based on media coverage, concluding that “defendants [need to] submit qualified public opinion surveys, other opinion testimony, or any other evidence demonstrating community bias caused by the media coverage”). For a discussion of surveys used in motions for change of venue, see Neal Miller, *Facts, Expert Facts, and Statistics: Descriptive and Experimental Research Methods in Litigation, Part II*, 40 Rutgers L. Rev. 467, 470–74 (1988); National Jury Project, *Jurywork: Systematic Techniques* (2d ed. 2008).

177. *McDonald’s Corp. v. McBagel’s, Inc.*, 649 F. Supp. 1268, 1278 (S.D.N.Y. 1986).

178. See *Thompson Med. Co. v. Pfizer Inc.*, 753 F.2d 208 (2d Cir. 1985); *Inc. Publ’g Corp. v. Manhattan Mag., Inc.*, 616 F. Supp. 370 (S.D.N.Y. 1985), *aff’d without op.*, 788 F.2d 3 (2d Cir. 1986).

In evaluating the sampling used in a telephone survey, the trier of fact should consider:

1. Whether (when prospective respondents are not business personnel) some form of random-digit dialing¹⁷⁹ was used instead of or to supplement telephone numbers obtained from telephone directories, because a high percentage of all residential telephone numbers in some areas may be unlisted;¹⁸⁰
2. What types of phones were included—that is, landlines, cell phones, or both. Over ninety-six percent of American adults live in households with cell phones, leading to a shift in sampling procedures that sometimes rely only on cell phones.¹⁸¹
3. Whether the sampling procedures required the interviewer to sample within the household or business, instead of allowing the interviewer to administer the survey to any qualified individual who answered the telephone;¹⁸² and
4. Whether interviewers were required to call back multiple times at several different times of the day and on different days to increase the likelihood of contacting individuals or businesses with different schedules and, as a result, to reduce nonresponse levels.¹⁸³

Telephone surveys that do not include these procedures may not provide precise measures of the characteristics of a representative sample of respondents, but may be adequate for providing rough approximations. The vulnerability of the survey depends on the information being gathered. More elaborate procedures are advisable for achieving a representative sample of respondents if the survey instrument requests information that is likely to differ for individuals with listed

179. “Random-digit” dialing provides coverage of households with both listed and unlisted telephone numbers by generating numbers at random from the sampling frame of all possible telephone numbers. James M. Lepkowski, *Telephone Sampling Methods in the United States*, in *Telephone Survey Methodology* 81–91 (Robert M. Groves et al. eds., 1988).

180. Studies comparing listed and unlisted household characteristics show some important differences. *Id.* at 76.

181. Between 2% and 3% of the U.S. population has only landline access. Stephen J. Blumberg & Julian V. Luke, *Wireless Substitution: Early Release of Estimates from the National Health Interview Survey, January–June 2020* (U.S. Dep’t of Health & Hum. Servs., Feb. 2021), at 4, <https://perma.cc/5TSM-V88J>; see also Courtney Kennedy et al., *Implications of Moving Public Opinion Surveys to a Single-Frame Cell-Phone Random-Digit-Dial Design*, 82 Pub. Op. Q. 279 (2018), <https://doi.org/10.1093/poq/nfy016> (“Analysis of more than 250 survey questions show[ed] that when landlines [were] excluded, estimates change[d] by less than one percentage point, on average.”).

182. This is a consideration only if the survey is sampling individuals. If the survey is seeking information on the household, more than one individual may be able to answer questions on behalf of the household.

183. This applies equally to in-person interviews.

telephone numbers versus individuals with unlisted telephone numbers, individuals rarely at home versus those usually at home, or groups who are more versus less likely to rely on landlines or cell phones.

The report submitted by a survey expert who conducts a telephone survey should specify:

- The procedures that were used to identify potential respondents, including both the procedures used to select the telephone numbers that were called and the procedures used to identify the qualified individual to question;
- The number of telephone numbers for which no contact was made; and
- The number of contacted potential respondents who refused to participate in the survey (and how many follow-up attempts were made).¹⁸⁴

Like computer-assisted personal interviewing (CAPI),¹⁸⁵ computer-assisted telephone interviewing (CATI) facilitates the administration and data entry of large-scale surveys.¹⁸⁶ A computer protocol may be used to generate and dial telephone numbers as well as to guide the interviewer.

Mail Questionnaires

In general, mail surveys tend to be substantially less costly than both in-person and telephone surveys.¹⁸⁷ Procedures that raise response rates for mail surveys include multiple mailings, highly personalized communications, prepaid return envelopes, incentives or gratuities, assurances of confidentiality, first-class outgoing postage, and follow-up reminders.¹⁸⁸

A mail survey will not produce a high rate of return unless the recruitment process begins with an accurate and up-to-date list of names and addresses for

184. Additional disclosure and reporting features applicable to surveys in general are described in the section titled “Completeness and Accuracy of All Relevant Information in the Survey Report” below.

185. See *infra* text accompanying note 205.

186. See Tourangeau et al., *supra* note 88, at 289; Saris, *supra* note 173.

187. See Chase H. Harrison, *Mail Surveys and Paper Questionnaires*, in *Handbook of Survey Research*, *supra* note 1, at 498, 499.

188. See, e.g., Richard J. Fox et al., *Mail Survey Response Rate: A Meta-Analysis of Selected Techniques for Inducing Response*, 52 *Pub. Op. Q.* 467, 482–84 (1988), <https://doi.org/10.1086/269125>; Kenneth D. Hopkins & Arlen R. Gullickson, *Response Rates in Survey Research: A Meta-Analysis of the Effects of Monetary Gratuities*, 61 *J. Experimental Educ.* 52, 54–57, 59 (1992), <https://doi.org/10.1080/00220973.1992.9943849>; Eleanor Singer et al., *Confidentiality Assurances and Response: A Quantitative Review of the Experimental Literature*, 59 *Pub. Op. Q.* 66, 71 (1995), <https://doi.org/10.1086/269458>; see generally Don A. Dillman et al., *Internet, Mail, and Mixed-Mode Surveys: The Tailored Design Method* (3d ed. 2009).

the target population. Even if the sampling frame is adequate, the sample may be unrepresentative if some individuals are more likely to respond than others (i.e., differential nonresponse). For example, if a survey targets a population that includes individuals with literacy problems, these individuals will tend to be underrepresented. Open-ended questions are generally of limited value on a mail survey because they depend entirely on the respondent to answer fully and do not provide the opportunity to probe or clarify unclear answers. Similarly, if eligibility to answer some questions depends on the respondent's answers to previous questions, such skip sequences may be difficult for some respondents to follow. Finally, because respondents complete mail surveys without supervision, survey personnel are unable to prevent respondents from discussing the questions and answers with others before completing the survey or to control the order in which respondents answer the questions. Although skilled design of questionnaire format, question order, and the appearance of the individual pages of a survey can minimize these problems, if it is crucial to have respondents answer questions in a particular order, a mail survey cannot be depended on to provide adequate data.

Internet Surveys

Over the past two decades there has been a massive increase in the use of internet surveys. One recent analysis found that the overwhelming majority of expert reports reviewed in the trademark and false advertising space were based on online surveys.¹⁸⁹ As of 2024, “an estimated” 96% of U.S. adults have access to the internet, including 99% of those between the ages of eighteen and forty-nine.¹⁹⁰ Although concerns remain about internet survey administration, and especially internet administration through online panels, many of those concerns can be mitigated through quality-control measures.

The benefits of internet administration are clear. The cost of conducting surveys online is a small fraction of the cost of hiring human interviewers to speak with potential respondents; the speed with which a researcher can find, screen, and survey hundreds of respondents online is far greater than in non-internet-based environments (e.g., via the mail); and there is less risk of interviewer bias and error when questions are presented by computer on a screen and respondents

189. Kugler & Henn, *supra* note 71, at 293–94.

190. Pew Rsch. Ctr., Internet, Broadband Fact Sheet (Nov. 2024), <https://www.pewinternet.org/fact-sheet/internet-broadband>. Of those age fifty to sixty-four, 98% had internet access. The sixty-five-plus group was lower at 90%. The same research shows that 79% of the U.S. population has broadband internet at home, while an additional 15% do not have broadband but do have internet access via smartphone. In 2011, only 79% of Americans had internet access. *Id.*

type their own responses.¹⁹¹ Further, internet surveys may be particularly appropriate when the goal of the researcher is to simulate everyday commercial conduct that now frequently occurs online.

Computer administration also allows careful randomization of survey items, branching survey logic, and the display of complicated visual and auditory stimuli. In addition, the structure permits the survey to remind, or even require, the respondent to answer a question before the next question is presented. A further advantage of computer-administered surveys over interviewer-administered surveys is that they eliminate interviewer error because the computer presents the questions and the respondent records their own answers.

One major question for internet surveys is the adequacy of the sampling approach. The evaluation of this will depend on the type of internet survey involved, because the samples used in web-based surveys vary in fundamental ways. At one extreme is the list-based web survey. This web-based survey is sent to a closed set of potential respondents drawn from a list that consists of the email addresses of the target individuals (e.g., all employees at a company where each employee has a known email address). Here there is no reason not to use the tools of probability sampling, described above.

At the other extreme is the self-selected web pseudosurvey in which web users in general, or those who happen to visit a particular website, are all invited to express their views on a topic and they participate simply by volunteering. Participants are very likely to self-select on the basis of the nature of the topic, substantially distorting the results. These self-selected pseudosurveys resemble reader polls published in magazines and do not meet standard criteria for legitimate surveys admissible in court.¹⁹² Occasionally, proponents of such polls tout the large number of respondents as evidence of the weight the results should be given, but the size of the sample cannot cure the likely participation bias in such voluntary polls.¹⁹³

Between these two extremes is a large category of web-based survey approaches that researchers have developed to address concerns about sampling bias and nonresponse error. Many of these use the professionally managed non-probability panels described above. These companies recruit diverse panels of respondents. Some respondents come from other well-traveled sites; others are

191. Reg Baker et al., *Research Synthesis: AAPOR Report on Online Panels*, 74 Pub. Op. Q. 711, 739 (2010), <https://doi.org/10.1093/poq/nfq048> (“Overall, the research reported here generally suggests higher data quality for computer administration than for oral administration.”).

192. See, e.g., *Merisant Co. v. McNeil Nutritionals, LLC*, 242 F.R.D. 315 (E.D. Pa. 2007) (report on results from AOL “instant poll” excluded).

193. See Mick P. Couper, *Web Surveys: A Review of Issues and Approaches*, 64 Pub. Op. Q. 464, 480–81 (2000), <https://doi.org/10.1086/318641> (a self-selected web survey conducted by the National Geographic Society through its website attracted 50,000 responses; a comparison of the Canadian respondents with data from the Canadian General Social Survey telephone survey conducted using random digit dialing showed marked differences on a variety of response measures).

pursued because their qualifications make them valuable for marketing studies. When a researcher contracts with a panel provider, the company can sift through its database to invite an appropriate mix of people to participate in a given survey. Those invited do not know the topic of the survey in advance (i.e., they are not opting in based on the topic). The final sample can be balanced to match desired demographics either through recruitment quotas or through response weighting.¹⁹⁴ The researcher using such data should obtain the procedures used by the vendor to ensure the quality of the data and make it available to the opposing party and the trier of fact. For example, what steps did the vendor take to minimize fraudulent respondents¹⁹⁵ and minimize duplicative respondents? Did the vendor themselves outsource the sampling?¹⁹⁶

Generally, a researcher conducting an internet survey must take active steps to ensure an honest and attentive sample. On the most technical level, use of CAPTCHA¹⁹⁷ questions can weed bots out from the sample; panel providers can audit their panels to verify participant demographics, or at least check that self-reported demographics are consistent over time,¹⁹⁸ and survey platforms, which host the actual survey questions, can use cookies and other means to prevent multiple submissions from the same person or computer.¹⁹⁹ As with many other issues in the online space, there is constant evolution in the fraud-detection domain. A researcher should ensure that they, and their panel provider, are using current best practices to detect fraudulent responses and be prepared to explain how these responses work and their known impact.

Moving beyond the technology and into the survey itself, the researcher should take further steps to ensure an attentive sample. In the trademark space, a

194. See, e.g., *Philip Morris USA, Inc. v. Otamedia Ltd.*, No. 02 Civ 7575 (GEL) (KNF), 2005 U.S. Dist. LEXIS 1259, at *10–11 (S.D.N.Y. Jan. 28, 2005).

195. Andrew M. Bell & Thomas Gift, *Fraud in Online Surveys: Evidence from a Nonprobability, Subpopulation Sample*, 10 J. Experimental Pol. Sci. 148–53 (2023), <https://doi.org/10.1017/XPS.2022.8>.

196. Peter K. Enns & Jake Rothschild, *Do You Know Where Your Survey Data Come From?*, Medium, May 2, 2022, <https://perma.cc/BD9T-SK25>.

197. CAPTCHA is an acronym for “Completely Automated Public Turing test to tell Computers and Humans Apart.” The respondent may be asked to decipher a set of distorted letters or complete a categorization task that would be difficult to automate.

198. Courtney Kennedy et al., *Pew Rsch. Ctr., Evaluating Online Nonprobability Surveys* 36 (May 2, 2016), <https://perma.cc/VJ4L-BTJF>:

Most panels are double opt-in, meaning potential panelists first enter an email address and then respond to an email sent from the panel provider in order to confirm the email account. Depending on the vendor, other quality control features include IP address validation and digital fingerprinting, which guards against a single person having multiple accounts in a given panel.

199. “The most common technique for identifying duplicate respondents is digital fingerprinting. Specific applications of this technique vary, but they all involve the capture of technical information about a respondent’s IP address, browser, software settings, and hardware configuration to construct a unique ID for that computer.” Baker et al., *supra* note 191, at 756 (emphasis omitted).

few checks are particularly common.²⁰⁰ One method of measuring attention in a traditional likelihood-of-confusion survey is to include a free-response question prompting participants to explain a prior answer. If the participant responds with gibberish or irrelevant responses, they are either not paying attention or not taking the survey seriously.²⁰¹ A similar check is possible in a Teflon survey testing whether a mark is generic.²⁰² There, a participant is asked whether each in a list of terms is a “brand name” or a “common name.” A participant who responds that all terms are brand names or all are common names, that is, who gives a straight-line response, should be excluded. They either do not understand the task or are not trying to do it. For surveys aimed at purchasers of a particular product, it is common to ask which, of a list of products, they have bought in the past X months. A researcher can include nonexistent products on those lists, screening out people who implausibly claim to have purchased them.

The science behind these efforts to ensure honest responding is ever-evolving. A researcher should employ some of the above mechanisms in any online survey, and should be prepared to justify their choices. A court should not treat the methods described here as a comprehensive checklist, however. No particular check is independently necessary or sufficient, but use of appropriate checks is required to ensure quality when the respondent is completing the survey online.

Mixed-Format Surveys

A final approach to data collection does not depend on a single mode, but instead involves a mixed-mode approach. By combining modes, the survey design may increase the likelihood that all sampled members of the target population will be contacted. For example, a person without a landline may be reached by mail or email. Similarly, response rates may be increased if members of the target population are more likely to respond to one mode of contact versus another. For example, a person unwilling to be interviewed by phone may respond to a written or email contact. If a mixed-mode approach is used, the questions and structure of the questionnaires are likely to differ across modes, and the expert should be prepared to address the potential impact of mode on the answers obtained.²⁰³

200. These checks are described in greater length in Kugler & Henn, *supra* note 71, at 300–06.

201. *Nat'l Fin. Partners Corp. v. Paycom Software, Inc.*, No. 14 C 7424, 2015 U.S. Dist. LEXIS 74700, at *25–26 (N.D. Ill. June 10, 2015) (questioning the results of a survey in part because the expert did not appear to have even read the free response answers, which included responses like “cool” and “LOL,” when the expert should have excluded nonresponsive answers).

202. See E. Deborah Jay, *Genericness Surveys in Trademark Disputes: Under the Gavel*, in *Trademark and Deceptive Advertising Surveys* 107, 113–16, 120–25 (Shari Diamond & Jerre Swann eds., 2d ed. 2022).

203. Don A Dillman & Benjamin L. Messer, *Mixed-Mode Surveys*, in *Handbook of Survey Research*, *supra* note 1, at 550, 553.

Surveys Involving Interviewers

Selection and Training of the Interviewers

When interviews are used to question survey respondents, the results are trustworthy only if “sound interview procedures were followed by competent interviewers.”²⁰⁴ Properly trained interviewers receive detailed written instructions on everything they are to say to respondents, any stimulus materials they are to use in the survey, and how they are to complete the interview form. These instructions should be made available to the opposing party and to the trier of fact. Interviewers should be instructed to record verbatim the respondent’s answers, to indicate explicitly whenever they repeat a question to the respondent, and to record any statements they make to the respondent or supplementary questions they ask.

Interviewers require training to ensure that they can follow directions in administering the survey questions and employ optimal interviewing techniques (e.g., pausing to give the respondent enough time to answer, resisting invitations to express their own beliefs or opinions, knowing when and how to use probes). Although procedures vary, there is evidence that interviewer performance suffers with less than a day of training in general interviewing skills and techniques for new interviewers.²⁰⁵ Additional training is needed when the interviewer is responsible for last-stage sampling (i.e., selecting the particular respondents to be interviewed in an unbiased fashion). Further, in-person interviews also should be conducted in situations without distractions and where others cannot overhear the answers. Failure to follow these dictates was one ground used by a court to reject as inadmissible a survey that purported to demonstrate consumer confusion.²⁰⁶

Some compromises may be accepted when surveys must be conducted swiftly. In trademark and deceptive advertising cases, the plaintiff’s usual request is for a preliminary injunction, because a delay means irreparable harm. Nonetheless, careful instruction and training of interviewers who administer the survey, as well as monitoring and validation to ensure quality control²⁰⁷ and complete disclosure of the methods used for all of the procedures followed, are crucial

204. *Toys “R” Us, Inc. v. Canarsie Kiddie Shop, Inc.*, 559 F. Supp. 1189, 1205 (E.D.N.Y. 1983). See also *Wisconsin v. Indivior Inc.* No. 16–5073, 2020 U.S. Dist. LEXIS 219949, at *58 (E.D. Pa. Nov. 24, 2020) (praising a survey expert for “training interviewers carefully on interviewing techniques and the subject matter of the survey”).

205. Fowler & Mangione, *supra* note 139, at 117; Nora Cate Schaeffer et al., *Interviewers and Interviewing*, in *Handbook of Survey Research*, *supra* note 1, at 437, 460.

206. *Toys “R” Us*, 559 F. Supp. at 1204 (some interviews apparently were conducted in a bowling alley; some interviewees waiting to be interviewed overheard the substance of the interview while they were waiting).

207. See section titled “Procedures Used to Ensure and Determine That the Survey Was Administered to Minimize Error and Bias” below.

elements that, if compromised, seriously undermine the trustworthiness of any survey.

Procedures Used to Ensure and Determine That the Survey Was Administered to Minimize Error and Bias

Three methods are used to ensure that the survey instrument was implemented in an unbiased fashion and according to instructions. The first, monitoring the interviews as they occur, is done most easily when telephone surveys are used, but can occur in the field via recordings (if respondents consent to it). Second, validation of interviews occurs when respondents in a sample are recontacted to ask whether the initial interviews took place and to determine whether the respondents were qualified to participate in the survey. The standard procedure for validation of in-person interviews is to telephone a random sample of about 10% to 15% of the respondents.²⁰⁸ This validation procedure does not determine whether the initial interview as a whole was conducted properly, but it warns interviewers that their work is being checked and can detect gross failures in the administration of the survey. In computer-assisted interviews, further validation information can be obtained from the timings that can be automatically recorded when an interview occurs.

A third way to verify that the interviews were conducted properly is to examine the work done by each individual interviewer. By reviewing the interviews and individual responses recorded by each interviewer and comparing patterns of response across interviewers, researchers can identify any response patterns or inconsistencies that warrant further investigation. When interviewers conduct the survey, the identity of the interviewer (or a unique code) should be included in the data provided to the opposing party and trier of fact.

When a survey is conducted at the request of a party for litigation rather than in the normal course of business, a heightened standard for validation checks may be appropriate. Thus, independent validation of a random sample of interviews by a third party rather than by the field service that conducted the interviews increases the trustworthiness of the survey results.²⁰⁹

208. See, e.g., *Davis v. S. Bell Tel. & Tel. Co.*, No. 89-2839-CIV-NESBITT, 1994 U.S. Dist. LEXIS 13257, at *16 (S.D. Fla. Feb. 1, 1994); *Nat'l Football League Props., Inc. v. N.J. Giants, Inc.*, 637 F. Supp. 507, 515 (D.N.J. 1986).

209. In *Rust Environment & Infrastructure, Inc. v. Teunissen*, 131 F.3d 1210, 1218 (7th Cir. 1997), the court criticized a survey in part because it “did not comport with accepted practice for independent validation of the results.”

Accuracy of Data Entry

Analyzing the results of a survey requires that the data obtained on each sampled element be recorded and often coded before the results can be tabulated and processed. Procedures for data entry should include checks for completeness, checks for reliability and accuracy, and rules for resolving inconsistencies. Accurate data entry is maximized when responses are verified by duplicate entry and comparison, and when data-entry personnel are unaware of the purposes of the survey.

The need for these checks and rules to control mistakes in data entry is particularly great when data collected from survey respondents are combined with data on those same respondents obtained from institutional databases or records (e.g., a survey of jurors is combined with court data on their jury service).²¹⁰ In such cases, both data entries and matches need to be closely scrutinized.

Disclosure and Reporting

Early Disclosure About the Survey Methodology and Results

Objections to the definition of the relevant population, the method of selecting the sample, and the wording of questions generally are raised for the first time when the results of the survey are presented. By that time, it is often too late to correct methodological deficiencies that could have been addressed in the planning stages of the survey. The plaintiff in a trademark case²¹¹ submitted a set of proposed survey questions to the trial judge, who ruled that the survey results would be admissible at trial while reserving the question of the weight the evidence would be given.²¹² The Seventh Circuit called this approach a commendable procedure and suggested that it would have been even more desirable if the parties had “attempt[ed] in good faith to agree upon the questions to be in such a survey.”²¹³

210. See generally Jan Van den Broeck et al., *Data Cleaning: Detecting, Diagnosing, and Editing Data Abnormalities*, 2 PLoS Med 2(10): e267, <https://doi.org/10.1371/journal.pmed.0020267>. See also Mary R. Rose & Marc A. Musick, *How Can You Tell if There Is a Crisis? Data and Measurement Challenges in Assessing Jury Representation*, 98 Chicago-Kent L. Rev. 35, 42 (2023).

211. *Union Carbide Corp. v. Ever-Ready, Inc.*, 392 F. Supp. 280, 291 (N.D. Ill. 1975), *rev'd*, 531 F.2d 366 (7th Cir. 1976).

212. Before trial, the presiding judge was appointed to the court of appeals, so the case was tried by another district court judge.

213. *Union Carbide*, 531 F.2d at 386. On one occasion, the Seventh Circuit recommended filing a motion in limine, asking the district court to determine the admissibility of a survey based on an examination of the survey questions and the results of a preliminary survey before the party undertakes the expense of conducting the actual survey. *Piper Aircraft Corp. v. Wag-Aero, Inc.*,

The *Manual for Complex Litigation, Second*, recommended that parties be required, “before conducting any poll, to provide other parties with an outline of the proposed form and methodology, including the particular questions that will be asked, the introductory statements or instructions that will be given, and other controls to be used in the interrogation process.”²¹⁴ The parties then were encouraged to attempt to resolve any methodological disagreements before the survey was conducted.²¹⁵ Although this passage in the second edition of the *Manual* has been cited with apparent approval,²¹⁶ the prior agreement that the *Manual* recommends has occurred rarely, and the *Manual for Complex Litigation, Fourth*, recommends, but does not advocate requiring, prior disclosure and discussion of survey plans.²¹⁷ As the *Manual* suggests, however, early disclosure can enable the parties to raise prompt objections that may permit corrective measures to be taken before a survey is completed.²¹⁸

Rule 26 of the Federal Rules of Civil Procedure requires extensive disclosure of the basis of opinions offered by testifying experts. However, Rule 26 does not produce disclosure of all survey materials, because parties are not obligated to disclose information about nontestifying experts. Parties considering whether to commission or use a survey for litigation are not obligated to present a survey that produces unfavorable results. Prior disclosure of a proposed survey instrument places the party that ultimately would prefer not to present the survey in the position of presenting damaging results or leaving the impression that the results are not being presented because they were unfavorable. Anticipating such a situation, parties do not decide whether an expert will testify until after the results of the survey are available.

Nonetheless, courts can encourage early disclosure and discussion even if they do not lead to agreement between the parties. In *McNeilab, Inc. v. American Home Products Corp.*,²¹⁹ Judge William C. Conner encouraged the parties to submit their survey plans for court approval to ensure their evidentiary value; the plaintiff did

741 F.2d 925, 929 (7th Cir. 1984). On another occasion, the parties jointly developed a survey administered by a neutral third-party survey firm. *Scott v. City of New York*, 591 F. Supp. 2d 554, 560 (S.D.N.Y. 2008) (survey design, including multiple pretests, negotiated with the help of the magistrate judge).

214. *Manual for Complex Litigation, Second* § 21.484 (1985).

215. *See id.*

216. *See, e.g., Nat'l Football League Props., Inc. v. N.J. Giants, Inc.*, 637 F. Supp. 507, 514 n.3 (D.N.J. 1986).

217. MCL 4th, *supra* note 41, § 11.493 (“including the specific questions that will be asked, the introductory statements or instructions that will be given, and other controls to be used in the interrogation process”).

218. *See id.*

219. 848 F.2d 34, 36 (2d Cir. 1988) (discussing with approval the actions of the district court). *See also* *Hubbard v. Midland Credit Mgmt.*, No. 1:05-cv-0216, 2009 U.S. Dist. LEXIS 13938 (S.D. Ind. Feb. 23, 2009) (court responded to plaintiff’s motions to approve survey methodology with a critique of the proposed methodology).

so and altered its research plan based on Judge Conner's recommendations. Parties can anticipate that changes consistent with a judicial suggestion are likely to increase the weight given to, or at least the prospects of admissibility of, the survey.²²⁰

Completeness and Accuracy of All Relevant Information in the Survey Report

The completeness of the survey report is one indicator of the trustworthiness of the survey and the professionalism of the expert who is presenting the results of the survey. The American Association for Public Opinion Research (AAPOR), a professional organization that brings together producers and users of survey data, offers standards for disclosure to be reported with any survey results.²²¹ The following list, which draws on the AAPOR guidelines, describes the elements that a survey report generally should include:

1. The purpose of the survey, its research sponsor, and those who conducted the research;
2. A definition of the target population;
3. A description of how the sample was recruited, including the method of selection (and how much of the population is covered); population quotas used, if any; respondent compensation, if any; the method (mode) of interview; the number of callbacks; respondent eligibility or screening criteria and method; and other pertinent information (e.g., the name of the vendor, if used);
4. A description of the results of sample implementation, including the number of
 - a. potential respondents contacted,
 - b. potential respondents not reached,
 - c. noneligibles,
 - d. refusals,
 - e. incomplete interviews or terminations, and
 - f. completed interviews and final sample size;
5. The dates of data collection;
6. The exact wording of the questions used, including a copy of the actual questionnaire (and screenshots of the questionnaire as viewed by

220. Larry C. Jones, *Developing and Using Survey Evidence in Trademark Litigation*, 19 Memphis St. U. L. Rev. 471, 481 (1989).

221. AAPOR also provides professional ethics guidelines as well as a transparency initiative where survey organizations can be certified as publicly disclosing their basic research methods and making them public, <https://perma.cc/4GPB-7QKK>.

respondents if administered electronically), interviewer instructions, and visual exhibits, with any cross-condition variations or randomizations marked;²²²

7. A description of any weighting or estimating procedures used;
8. A description of data-processing procedures (e.g., validity checks such as measures of the time respondents took to complete the survey, any data-imputation procedures, criteria for removing any responses);
9. Estimates of the sampling error, where appropriate (i.e., in probability samples);
10. Statistical tables clearly labeled and identified regarding the source of the data, including the number of raw cases forming the base for each table, row, or column;
11. Copies of interviewer instructions, validation results, and code books, including coding instructions for free-response data;²²³ and
12. Acknowledgment of limitations to the survey.

As a general rule, any research should provide such information or explain why doing so is not possible. Additional information to include in the survey report may depend on the nature of sampling design. For example, reported response rates along with the exact time each interview occurred may assist in evaluating the likelihood that nonresponses biased the results. In a survey designed to assess the duration of employee pre-shift activities, workers were approached as they entered the workplace; records were not kept on refusal rates or the timing of participation in the study. Thus, it was impossible to rule out the plausible hypothesis that individuals who arrived early for their shift with more time to spend on pre-shift activities were more likely to participate in the study.²²⁴

222. The questionnaire itself can often reveal important sources of bias. See *Marria v. Broadus*, 200 F. Supp. 2d 280, 289 (S.D.N.Y. 2002) (court excluded survey sent to prison administrators based on questionnaire that began, “We need your help. We are helping to defend the NYS Department of Correctional Service in a case that involves their policy on intercepting Five-Percenter literature. Your answers to the following questions will be helpful in preparing a defense.”).

223. Failure to supply this information substantially impairs a court’s ability to evaluate a survey. *In re Prudential Ins. Co. of Am. Sales Pracs. Litig.*, 962 F. Supp. 450, 532 (D.N.J. 1997) (citing the first edition of this manual). *But see Fla. Bar v. Went for It, Inc.*, 515 U.S. 618, 626–28 (1995), in which a majority of the Supreme Court relied on a summary of results prepared by the Florida Bar from a consumer survey purporting to show consumer objections to attorney solicitation by mail. In a strong dissent, Justice Kennedy, joined by three other justices, found the survey inadequate based on the document available to the court, pointing out that the summary included “no actual surveys, few indications of sample size or selection procedures, no explanations of methodology, and no discussion of excluded results . . . no description of the statistical universe or scientific framework that permits any productive use of the information the so-called Summary of Record contains.” *Id.* at 640 (Kennedy, J., dissenting).

224. See *Chavez v. IBP, Inc.*, No. CT-01-5093-RHW, 2004 U.S. Dist. LEXIS 28837 (E.D. Wash. Aug. 18, 2004).

Researchers should also ask, and disclose, whether respondents have previously participated in similar prior surveys. Many nonprobability samples come from vendors with online panels where respondents participate in multiple surveys over time. There also are vendors who have established analogous online panels with probability samples.²²⁵ Some evidence suggests that repeat respondents from these panels differ from fresh sample draws.²²⁶ Vendors often do not provide information on prior participation—something about which a researcher can inquire prior to using a given vendor. Even if not, researchers can, on their own, attempt to identify “professional respondents” and consider whether they might bias inferences (e.g., by measuring prior participation and evaluating its effect on outcomes).²²⁷ If the expert asks participants whether they have participated in previous studies on the same topic, or obtains that information from other sources, the results obtained from the experienced and naive participants can be compared. Currently, many experts take the sensible approach of simply excluding participants who report having completed similar prior surveys. Regardless, researchers should always disclose prior participation and its possible effects or be explicit that such information does not exist and explain why it was not obtained.²²⁸

When the presenting party has access to the raw data from a survey, that data should generally be made available to opposing counsel upon request. This dataset should include responses from all participants whose data were ultimately used in the original expert’s report and also all participants whose data were excluded by the original expert for reasons of quality after the completion of the survey, with an indication of why they were excluded (i.e., data-processing procedures).²²⁹ If a researcher wishes to exclude unusually fast and slow responses, they should disclose this in the report and provide the data from those respondents as well.²³⁰ This permits an opposing expert to rerun statistical analyses and determine if any debatable analytic choices made by the original expert had substantial effects on the report’s conclusions.

225. Online Panel Research: A Data Quality Perspective (Mario Callegaro et al, eds., 2014).

226. Andrew Halpern-Manners & John Robert Warren, *Panel Conditioning in Longitudinal Studies: Evidence From Labor Force Items in the Current Population Survey*, 49 *Demography* 1499 (2012), <https://doi.org/10.1007/s13524-012-0124-x>.

227. D. Sunshine Hillygus et al., *Professional Respondents in Nonprobability Online Panels*, in *Online Panel Research: A Data Quality Perspective* 219–37 (2014).

228. K.H. Jamieson et al., *Protecting the Integrity of Survey Research*, 2 *PNAS Nexus* 1 (2023), <https://doi.org/10.1093/pnasnexus/pgad049>. More generally, these authors supplement AAPOR’s code by suggesting researchers should provide details on question order, details on respondent attrition (i.e., dropping out of the survey), and so on.

229. For example, an expert may exclude respondents who responded with gibberish to free-response items or who completed the survey much more quickly than did other participants.

230. The issue of completion time-based exclusions is explored at greater length in Kugler & Henn, *supra* note 71, at 305–06. At present, there is no agreed-upon standard for identifying “too fast” responses; most researchers appear to be using rules of thumb (for example, excluding those who took less than one third the median time). *Id.*

The public (and, presumably, also judges) sometimes express concern that surveys are not sufficiently reliable, generally citing polling accuracy in recent elections.²³¹ Election pollsters have the unenviable task of modeling a changing electorate's intended behavior in a circumstance where it matters a great deal whether the correct answer is 49% support or 51% support. Other surveys do not require this level of precision relative to a specified standard (i.e., greater than 50%). Rather, the relevant estimate the survey is seeking to reflect is whether approximately 20% of people (vs. 40%, 60%, or even 5%) in the target market are confused by a given advertisement. Thus, the accuracy of a survey estimate should be assessed in light of the degree of precision needed for the context.

Though transparency is very important, some information must be removed from data files before they are shared beyond the survey expert to protect participant confidentiality. All identifying information, such as the respondent's name, address, and telephone number, should be omitted. In the case of internet surveys, it is appropriate to also remove the participant's panel ID number, IP address, and precise geolocation information; these might all be linked to the participant's identity.²³² Keeping the survey duration and survey start time in the file may aid analysis, however, and will not compromise participant anonymity in most cases.

Greater efforts to promote participant anonymity are appropriate in cases where the population surveyed is small or otherwise susceptible to easy reidentification. If a survey is conducted of a workplace, for example, it may be that the total population numbers only in the hundreds. The identity of a respondent of any given age, ethnicity, and gender may therefore be narrowed down to one of only a few possibilities.²³³ Much greater care should be taken to protect participant anonymity in such cases. Possibilities include omitting data fields that the opposing expert does not expect to use in their own analyses, or restricting access

231. See, e.g., Scott Keeter et al., Pew Rsch. Ctr., *What 2020's Election Poll Errors Tell Us About the Accuracy of Issue Polling* (Mar. 2, 2021), <https://perma.cc/LDF9-KUWK> (commenting on this concern and evaluating it). For a discussion of the accuracy of 2020 polling, see American Association of Public Opinion Research, *Task Force on 2020 Pre-Election Polling: An Evaluation of the 2020 General Election Polls*, <https://perma.cc/64K9-GVDJ> (not reaching firm conclusions, but suggesting that this may have been due to greater nonresponse among Trump voters and difficulty in accounting for new voters in screening).

232. IP addresses are sometimes used to detect whether individuals took the survey more than once. If so, shared data should indicate which respondents shared an IP address but should generally not include the IP address itself.

233. This problem also arises in the case of privacy statutes such as HIPAA and FERPA, where deidentified data can be shared but identifiable data is tightly restricted. The Department of Education reviews a number of questions that may be relevant to a school or workplace survey in their FERPA guidance, <https://perma.cc/CCD3-47Y8> (updated May 2013).

to the data to ensure that it is not shared beyond the opposing expert—and particularly not with the ultimate client.²³⁴

The respondents questioned in a survey generally do not testify in legal proceedings and are unavailable for cross-examination. Indeed, one of the advantages of a survey is that it avoids a repetitious and unrepresentative parade of witnesses. To verify that interviews occurred with qualified respondents, standard survey practice includes validation procedures,²³⁵ the results of which should be included in the survey report.

Conflicts may arise when an opposing party asks for survey respondents' names and addresses so that they can re-interview some respondents. The party introducing the survey or the survey organization that conducted the research generally resists supplying such information.²³⁶ Professional surveyors as a rule promise confidentiality to increase participation rates and encourage candid responses although, to the extent that identifying information is collected, such promises may not effectively prevent a lawful inquiry. Because failure to extend confidentiality may bias both the willingness of potential respondents to participate in a survey and their responses, the professional standards for survey researchers generally prohibit disclosure of respondents' identities. "The use of survey results in a legal proceeding does not relieve the Survey Research Organization of its ethical obligation to maintain in confidence all Respondent-identifiable information or lessen the importance of Respondent anonymity."²³⁷ Although no surveyor–respondent privilege currently is recognized, the need for surveys and the availability of other means to examine and ensure their trustworthiness argue for deference to legitimate claims for confidentiality in order to avoid seriously compromising the ability of surveys to produce accurate information.²³⁸ In

234. In the case of a workplace survey, the client might well be the current employer of the survey participant.

235. See section titled "Procedures Used to Ensure and Determine That the Survey Was Administered to Minimize Error and Bias" above.

236. See, e.g., *Alpo Petfoods, Inc. v. Ralston Purina Co.*, 720 F. Supp. 194 (D.D.C. 1989), *aff'd in part and vacated in part*, 913 F.2d 958 (D.C. Cir. 1990).

237. CASRO, *supra* note 41, § I.A.3f; Am. Ass'n for Pub. Op. Rsch., AAPOR Code of Professional Ethics and Practices (revised Apr. 2021), <https://perma.cc/DBC8-LTWJ>.

238. *United States v. Dentsply Int'l, Inc.*, No. 99–5 MMS, 2000 U.S. Dist. LEXIS 6994, at *23 (D. Del. May 10, 2000) (Fed. R. Civ. P. 26(a)(1) does not require party to produce the identities of individual survey respondents); *In re Litton Indus., Inc.*, No. 9123, 1979 FTC LEXIS 311, at *13 & n.12 (June 19, 1979) (Order Concerning the Identification of Individual Survey-Respondents with Their Questionnaires) (citing Frederick H. Boness & John F. Cordes, *The Researcher–Subject Relationship: The Need for Protection and a Model Statute*, 62 Geo. L.J. 243, 253 (1973)); see also *Applera Corp. v. MJ Rsch., Inc.*, 389 F. Supp. 2d 344, 350 (D. Conn. 2005) (denying access to names of survey respondents); *Lampshire v. Procter & Gamble Co.*, 94 F.R.D. 58, 60 (N.D. Ga. 1982) (defendant denied access to personal identifying information about women involved in studies by the Centers for Disease Control based on Fed. R. Civ. P. 26(c) giving court the authority to enter

general, the better approach is for the opposing party to conduct their own survey rather than re-interrogate the participants in the previous one.

Concluding Observations

As this chapter has shown, judges have a variety of factors to consider when determining whether a survey should be admitted and how much weight it should be given. The *Manual for Complex Litigation, Fourth (MCL4)*, published in 2004, suggested a set of relevant factors for judges to evaluate.²³⁹ In the past twenty years since the *MCL4* was published, survey methods have evolved (e.g., with the growth of internet and other computer technology, and recognition of the importance of survey-experimental methodology for causal inference), but these factors remain important. Thus, we close by presenting an updated list that clarifies and builds on the list of factors presented in the *MCL4*.²⁴⁰

Relevant factors include whether:

- the population was properly chosen and defined;
- the sampling frame and the sample itself were representative of that population;
- the data and details on data collection were fully and accurately reported;
- the survey questions asked were clear and not leading;
- the survey was conducted by qualified persons following proper procedures;
- the data were analyzed following accepted statistical principles;
- statements about causal relationships were supported by sufficient evidence about causality; and
- the results were presented in accordance with statistical standards.²⁴¹

“any order which justice requires to protect a party or persons from annoyance, embarrassment, oppression, or undue burden or expense”) (citation omitted).

239. MCL 4th, *supra* note 41, § 11.493.

240. The *Manual for Complex Litigation* distinguished between factors to be considered regarding admissibility and factors to be considered in assigning weight. As all of the factors can affect either admissibility or weight, depending on their quality, we have combined them in one set.

241. These include, where appropriate, information on sampling error and confidence intervals.

Glossary of Terms

The following terms and definitions were adapted from a variety of sources, including *Handbook of Survey Research* (Peter H. Rossi et al. eds., 1st ed. 1983; Peter V. Marsden & James D. Wright eds., 2d ed. 2010); *Measurement Errors in Surveys* (Paul P. Biemer et al. eds., 1991); Willem E. Saris, *Computer-Assisted Interviewing* (1991); Seymour Sudman, *Applied Sampling* (1976).

branching. A questionnaire structure that uses the answers to earlier questions to determine which set of additional questions should be asked (e.g., citizens who report having served as jurors on a criminal case are asked different questions about their experiences than citizens who report having served as jurors on a civil case).

CAI (computer-assisted interviewing). A method of conducting interviews in which an interviewer asks questions and records the respondent's answers by following a computer-generated protocol.

CAPI (computer-assisted personal interviewing). A method of conducting face-to-face interviews in which an interviewer asks questions and records the respondent's answers by following a computer-generated protocol.

CATI (computer-assisted telephone interviewing). A method of conducting telephone interviews in which an interviewer asks questions and records the respondent's answers by following a computer-generated protocol.

closed-ended question. A question that provides the respondent with a list of choices and asks the respondent to choose from among them.

cluster sampling. A sampling technique allowing for the selection of sample elements in groups or clusters, rather than on an individual basis; it may significantly reduce field costs and may increase sampling error if elements in the same cluster are more similar to one another than are elements in different clusters.

confidence interval. An indication of the probable range of error associated with a sample value obtained from a probability sample.

conjoint survey. Survey-experiment designed to identify and estimate the causal effects of many treatment components simultaneously by using an orthogonal, fractional factorial design.

context effect. When a previous question influences the way the respondent perceives and answers a later question.

convenience sample. A sample of elements selected because they were readily available.

coverage error. Any inconsistencies between the sampling frame and the target population.

double-blind research. Research in which the respondent and the interviewer are not given information that will alert them to the anticipated or preferred pattern of response.

full-filter question. A question asked of respondents to screen out those who do not have an opinion on the issue under investigation before asking them the question proper.

mall intercept survey. A survey conducted in a mall or shopping center in which potential respondents are approached by a recruiter (intercepted) and invited to participate in the survey.

margin of error. An indication of the likely precision of an estimate from a probability sample; used to compute a confidence interval.

multistage sampling design. A sampling design in which sampling takes place in several stages, beginning with larger units (e.g., cities) and then proceeding with smaller units (e.g., households or individuals within these units).

nonprobability sample. Any sample that does not qualify as a probability sample.

open-ended question. A question that requires the respondent to formulate their own response.

order effect. A tendency of respondents to choose an item based in part on the order of response alternatives on the questionnaire (see primacy effect and recency effect).

parameter. See population value.

pilot test. A small field test replicating the field procedures planned for the full-scale survey; although the terms *pilot test* and *pretest* are sometimes used interchangeably, a pretest tests the questionnaire, whereas a pilot test generally tests proposed collection procedures as well.

population. The totality of elements (objects, individuals, or other social units) that have some common property of interest; the target population is the collection of elements that the researcher would like to study. Also, universe.

population value, population parameter. The actual value of some characteristic in the population (e.g., the average age); the population value is estimated by taking a random sample from the population and computing the corresponding sample value.

pretest. A small preliminary test of a survey questionnaire. See pilot test.

primacy effect. A tendency of respondents to choose early items from a list of choices; the opposite of a recency effect.

probability sample. A type of sample selected so that every element in the population has a known nonzero probability of being included in the sample; a simple random sample is a probability sample.

probe. A follow-up question that an interviewer asks to obtain a more complete answer from a respondent (e.g., “Anything else?” “What kind of medical problem do you mean?”).

quasi-filter question. A question that offers a “don’t know” or “no opinion” option to respondents as part of a set of response alternatives; used to screen out respondents who may not have an opinion on the issue under investigation.

random sample. See probability sample.

recency effect. A tendency of respondents to choose later items from a list of choices; the opposite of a primacy effect.

sample. A subset of a population or universe selected so as to yield information about the population as a whole.

sampling error. The estimated size of the difference between the result obtained from a sample study and the result that would be obtained by attempting a complete study of all units in the sampling frame from which the sample was selected in the same manner and with the same care.

sampling frame. The source or sources from which the objects, individuals, or other social units in a sample are drawn.

secondary meaning. A descriptive term that becomes protectable as a trademark if it signifies to the purchasing public that the product comes from a single producer or source.

simple random sample. The most basic type of probability sample; each unit in the population has an equal probability of being in the sample, and all possible samples of a given size are equally likely to be selected.

skip pattern, skip sequence. A sequence of questions in which some should not be asked (should be skipped) based on the respondent’s answer to a previous question (e.g., if the respondent indicates that he does not own a car, he should not be asked what brand of car he owns).

stratified sampling. A sampling technique in which the researcher subdivides the population into mutually exclusive and exhaustive subpopulations, or strata; within these strata, separate samples are selected. Results can be combined to form overall population estimates or used to report separate within-stratum estimates.

survey-experiment. A survey with randomly assigned control and treatment groups, enabling the researcher to test a causal proposition.

survey population. See population.

trade dress. A distinctive and nonfunctional design of a package or product protected under state unfair competition law and the federal Lanham Act § 43(a), 15 U.S.C. § 1125(a) (1946) (amended 1992).

universe. See population.

References on Survey Research

- Jean M. Converse & Stanley Presser, *Survey Questions: Handcrafting the Standardized Questionnaire* (1986).
- Mick P. Couper, *Designing Effective Web Surveys* (2008).
- Matthew DeBell, *Computation of Survey Weights*, in *The Palgrave Handbook of Survey Research*, 519 (David L. Vannette & Jon A. Krosnick eds., 2018).
- Shari S. Diamond, *Control Foundations: Rationale and Approaches*, in *Trademark and Deceptive Advertising Surveys: Law, Science and Design* 239 (Shari S. Diamond and & Jerre Swann, eds., 2d ed. 2022).
- Don A. Dillman, Jolene Smyth & Leah M. Christian, *Internet, Mail and Mixed-Mode Surveys: The Tailored Design Method* (3d ed. 2009).
- James N. Druckman, *Experimental Thinking: A Primer on Social Science Experiments* (2022).
- Experimental Methods in Survey Research: Techniques that Combine Random Sampling and Random Assignment* (Paul J. Lavrakas, Michael W. Traugott, Courtney Kennedy, Allyson L. Holbrook, Edith D. de Leeuw & Brady T. West eds., 2019).
- Arlene Fink, *How to Conduct Surveys: A Step-By-Step Guide* (4th ed. 2009).
- Robert M. Groves, Floyd J. Fowler, Jr., Mick P. Couper, James M. Lepkowski, Eleanor Singer & Roger Tourangeau, *Survey Methodology* (2d ed. 2009).
- Handbook of Survey Research* (Peter V. Marsden & James D. Wright eds., 2d ed. 2010).
- Matthew B. Kugler & R. Charles Henn, *Internet Surveys in Trademark Cases: Benefits, Challenges, and Solutions*, in *Trademark and Deceptive Advertising Surveys: Law, Science and Design*, 293 (Shari S. Diamond and Jerre B. Swann, eds., 2d ed. 2022).
- Sharon Lohr, *Sampling: Design and Analysis* (2d ed. 2010).
- Measurement Errors in Surveys* (Paul P. Biemer, Robert M. Groves, Lars E. Lyberg, Nancy A. Mathiowetz & Seymour Sudman eds., 2004).
- Online Panel Research: A Data Quality Perspective* (Mario Callegaro, Reg Baker, Jelke Bethlehem, Anja S. Göritz, Jon A. Krosnick & Paul J. Lavrakas eds., 2014).
- The Palgrave Handbook of Survey Research* (David L. Vannette & Jon A. Krosnick eds., 2018).
- Questions About Questions: Inquiries into the Cognitive Bases of Surveys* (Judith M. Tanur ed., 1992).
- Howard Schuman & Stanley Presser, *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording and Context* (1981).
- William Shadish, Thomas D. Cook & Donald T. Campbell, *Experimental and Quasi-Experimental Designs for Generalized Causal Inferences* (2002).

Reference Manual on Scientific Evidence

- Monroe G. Sirken, Douglas J. Herrmann, Susan Schechter, Norbert Schwarz, Judith M. Tanur & Roger Tourangeau, *Cognition and Survey Research* (1999).
- Seymour Sudman, *Applied Sampling* (1976).
- Survey Nonresponse* (Robert M. Groves, Don A. Dillman, John L. Eltinge & Roderick J. A. Little eds., 2002).
- Telephone Survey Methodology* (Robert M. Groves, Paul P. Biemer, Lars E. Lyberg, James T. Massey & William L. Nicholls eds., 1988).
- Roger Tourangeau, Lance J. Rips & Kenneth Rasinski, *The Psychology of Survey Response* (2000).
- Trademark and Deceptive Advertising Surveys: Law, Science and Design* (Shari S. Diamond & Jerre B. Swann eds., 2d ed. 2021).

Reference Guide on Estimation of Economic Damages

MARK A. ALLEN, CARLOS BRAIN, AND FILIPE LACERDA

Mark A. Allen, J.D., is Principal at Cornerstone Research.

Carlos Brain, M.B.A., Sc.M., is Vice President at Cornerstone Research.

Filipe Lacerda, Ph.D., M.B.A., Finance, is Vice President at Cornerstone Research.

The views expressed herein are solely those of the authors, who are responsible for the content, and do not necessarily represent the views of Cornerstone Research.

CONTENTS

Introduction, 752

Damages Experts' Qualifications, 752

The Standard General Approach to Quantification of Economic Damages, 754

General Principles, 754

Isolating the Effect of the Harmful Act, 755

General Categories of Damages Measures, 757

Basic Issues Arising in Damages Estimation, 759

Losses Caused, 759

Hypothesizing Legitimate Conduct of the Defendant in Analyzing the Plaintiff's Economic Position But For the Harmful Event, 761

Considering All the Differences in the Plaintiff's Situation in the But-For Scenario Versus Assuming That Many Aspects Would Be the Same as in Actuality, 762

Valuation Issues in the Analysis of Economic Damages, 764

Issues Arising When the Potential Impact of the Alleged Harmful Act

Is Analyzed Directly on a Particular Date, 765

Issues Arising in Connection with the Use of the Market Price of the Product or Asset at Issue, 765

Issues Arising in Connection with the Use of the Market Price of Ostensibly Similar Products or Assets, 767

Identifying products or assets that are ostensibly similar to the allegedly affected product or asset, 768

Adjusting the market price of ostensibly similar products or assets, 768

Issues Arising in Connection with the Use of Models to Simulate the But-For Price, 770

Realistic modeling of preferences, incentives, and constraints of buyers and sellers, 771

- Modeling the impact of the alleged misconduct on demand and supply, 772
- Accounting for Market Frictions, 773
- Issues Arising in Connection with the Use of Repair or Abatement Costs, 775
- Issues Arising When the Potential Impact of the Alleged Harmful Act Spans Multiple Dates and Damages Are Analyzed Indirectly on a Single Date, 776
 - Estimating the Period-by-Period Direct Impact of the Alleged Harmful Act, 776
 - Disputes about how the defendant's actions would have differed had the alleged harmful act not occurred, 776
 - Disputes about how the plaintiff's actions would have differed had the alleged harmful act not occurred, 777
 - Disputes about whether all economic consequences of the alleged harmful act are being accounted for, 778
 - Disputes about the use of expected outcomes to estimate the impact of the alleged harmful act, 780
 - Disputes about the economic situation in future periods, 781
 - Disputes about the consideration of non-cash effects of the alleged harmful act, 782
 - Valuing the Period-by-Period Impact of the Alleged Harmful Act as of a Single Valuation Date, 783
 - Disagreements about the discount rate to value the impact of the alleged harmful act in periods after the valuation date, 783
 - Estimating the time value of money, 784
 - Estimating the risk premium, 785
 - Disagreements about the capitalization rate to value the impact of the alleged harmful act in periods before the valuation date, 786
 - Disagreements about what valuation date to use, 787
- Other Issues Arising in General in Damages Measurement, 788
 - Data Validity, 788
 - Criteria for Determining the Validity of Data, 788
 - Quantitative Methods for Validation, 789
 - Handling of Missing Data, 790
 - Prejudgment Interest, 791
 - Accounting for Income Taxes, 792
 - Damages with Multiple Challenged Acts: Disaggregation, 794
 - Disputes About Whether the Plaintiff Is Entitled to All the Damages, 796
- Limitations on Damages, 797
 - Uncertainty and Speculation in Damages Estimates, 797
 - Damages That Are Too Remote, 798
 - Mitigation of Losses, 799
 - Damages That May Exceed the Cost of Avoidance, 800

Reference Guide on Estimation of Economic Damages

The Effect of a Liquidated Damages Clause, 801
Damages in Class Actions, 802
Class Certification, 802
Class-Wide Damages, 804
Damages of Individual Class Members, 805
Damages as a Basis to Evaluate the Fairness of a Proposed Settlement, 805
Damages Issues in Selected Types of Cases, 806
Estimation of Economic Damages in Consumer Product Liability and Misrepresentation Cases, 806
Defining the Plaintiff's Economic Position in the Actual and But-For Worlds, 806
Determining the But-For Price, 807
Conjoint Analysis and Supply-Side Considerations, 808
Estimation of Economic Damages in Securities Cases, 810
Inflation and Losses Caused, 811
Key Areas of Disagreement in the Economic Analysis of Inflation and Losses Caused, 813
Event study model, 813
Analysis of "confounding" information, 815
Measuring inflation at the time of purchase, 815
Estimation of Economic Damages in Antitrust Cases, 816
Threshold Issues, 816
Quantifying Damages in an Antitrust Matter, 817
Overcharge Damages, 817
Before-and-after model of overcharge damages, 818
Difference-in-differences model of overcharge damages, 820
Lost-Profits Damages, 821
Estimation of Economic Damages from Loss of Personal Income, 823
Estimation of Losses Over a Person's Lifetime, 823
Calculation of Fringe Benefits, 824
Medical insurance benefits, 824
Retirement benefits, 825
Defined benefit plan, 825
Defined contribution plan, 826
Wrongful Death, 826
Shortened Life Expectancy, 827
Acknowledgments, 827
Glossary of Terms, 828

FIGURE

1. Damages studies generally adopt this framework. We therefore use it as the basic model for this reference guide, 755

TABLE

1. Calculations of Prejudgment Interest (in Dollars), 791

Introduction

This reference guide identifies areas of dispute that arise when economic losses are at issue in legal proceedings. We focus on explaining the issues in these disputes rather than taking positions on their proper resolution. We discuss the application of economic analysis within established legal frameworks for damages, cover topics in economics that arise in measuring damages, and, where useful, provide citations to cases that illustrate the principles and techniques discussed in the text.

We begin with a brief discussion of qualifications for damages experts. We then set forth the standard general approach to quantification of economic damages, with particular focus on translating the legal theory of the harmful event into an analysis of the economic impact of that event and choosing the measure of economic damages. Next, we consider disputes that may arise between the parties regarding damages estimation. Such disputes may concern basic issues (e.g., proper damages measure and losses caused), valuation issues (e.g., damages arising on single or multiple dates, and the use of market or other data to estimate damages), other issues (e.g., data validity, prejudgment interest, disaggregation), and limitations on damages. We also discuss damages in class actions as well as the application of damages principles to specific types of cases.

We use brief examples to illustrate the disputes that can arise and provide intuition for relevant economic issues. These examples are not full case descriptions; they are deliberately stylized, attempting to capture the types of disagreements about damages that arise in practical experience, although they are purely hypothetical. In many examples, the dispute involves factual and economic as well as legal issues. We do not try to resolve the disputes in these examples; hopefully the examples will help clarify the legal and factual disputes that need to be resolved before or at trial.

Damages Experts' Qualifications

Experts who quantify damages come from a variety of backgrounds. The expert should be trained and experienced in quantitative analysis. For economists, the common qualification is a Ph.D. degree. Damages experts with business or accounting backgrounds often have M.B.A. or other advanced degrees, or C.P.A. credentials. Both the method of damages estimation used and the substance of the damages claim dictate the specific areas of specialization the expert should have. In some cases, participation in original research and authorship of professional publications may add to the qualifications of an expert. However, relevant research and publications are not likely to be on the topic of damages measurement per se, but rather on topics and methods encountered in damages analysis.

Professional training and experience in areas relevant to the substance of the damages claim may be required. For example, in antitrust, a background in industrial organization (i.e., competition economics) may be helpful; in securities damages, a background in finance may assist the expert; and in the case of lost earnings, an expert may benefit from training in labor economics.

An analysis by even the most qualified expert may face a motion to exclude under Federal Rule of Evidence 702 as interpreted by the *Daubert* line of cases.¹ Under Rule 702, an expert witness who is “qualified . . . by knowledge, skill, experience, training, or education,” is allowed to testify when

- (a) the expert’s scientific, technical, or other specialized knowledge will help the trier of fact to understand the evidence or to determine a fact in issue; (b) the testimony is based on sufficient facts or data; (c) the testimony is the product of reliable principles and methods; and (d) the expert has reliably applied the principles and methods to the facts of the case.

These criteria, which are intended to exclude testimony that is irrelevant or is based on untested and unreliable theories, are often applied to damages analyses. Examples include the following:

- A proposed damages expert may be qualified generally to opine on damages issues but may lack the qualifications necessary to offer an opinion in the context of the particular matter at hand. For example, a labor economist may lack the proper education, training, and experience to opine on damages in a securities matter.
- A proposed expert may have used a methodology that is incorrect as a matter of economics; the results may have been based on insufficient data (e.g., too few observations); or a theoretically sound analysis may not have been properly applied, even if large amounts of data were used in the analysis.
- All or part of a damages analysis may no longer be relevant if certain legal claims have been dismissed, or expert opinion on damages simply may not be needed at all, such as where damages may be determined using a simple mathematical formula.

Even if a motion to exclude fails, it can be an effective way for a party to probe an opposing expert’s damages analysis prior to trial. For this reason, among others, motions to exclude proposed expert damages testimony have become relatively commonplace.

1. *Daubert v. Merrell Dow Pharms., Inc.*, 509 U.S. 579 (1993); *Kumho Tire Co. v. Carmichael*, 526 U.S. 137 (1999). For a discussion of emerging standards of scientific evidence, see Liesa L. Richter & Daniel J. Capra, *The Admissibility of Expert Testimony*, in this manual.

The Standard General Approach to Quantification of Economic Damages

In this section, we review the general principles of the standard approach to the estimation of economic damages and basic issues that arise in such estimation. For each principle, there are several areas of potential dispute. The sequence of issues discussed here is intended to identify most of the areas of disagreement between the damages analyses of opposing parties with respect to these principles.

General Principles

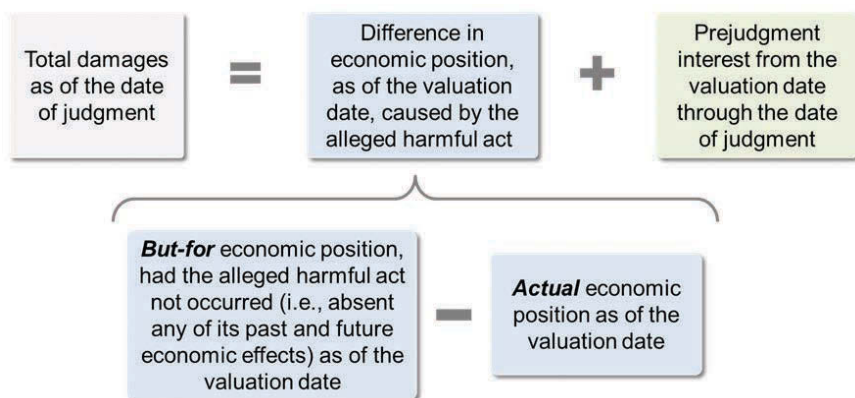
We begin by setting forth the standard general approach to damages quantification, with particular focus on defining the harmful event and the alternative, or but-for, scenario. In most cases, the goal of measurement of economic damages is to estimate the plaintiff's loss of economic value caused by the defendant's harmful act. The loss of value may arise from a single event, such as overpayment for a consumer product, or it may take the form of a reduced stream of profits or earnings over time. The losses are net of any costs avoided because of the harmful act.

In principle, then, economic damages are the difference between (1) the plaintiff's economic position assuming that the alleged harmful act had not occurred, and (2) the plaintiff's actual economic position as a result of the harmful act.² The former considers the plaintiff's economic position absent any past and future effects of the harmful act. We refer to this as the plaintiff's but-for economic position in the but-for world. The latter considers the plaintiff's actual economic position given the occurrence of the harmful act. We refer to this as the plaintiff's actual economic position in the actual world.

Figure 1 illustrates the structure of the standard damages study. We begin by estimating the plaintiff's losses as of the appropriate valuation date. The appropriate valuation date may vary depending on the nature of the loss and the plaintiff's theory of recovery (e.g., breach of contract, tort, fraud, etc.). Losses are the difference between the plaintiff's but-for economic position and the plaintiff's actual economic position as of the valuation date. Note that the plaintiff's but-for economic position may require estimating past losses as well as future losses, again

2. Damages are sometimes measured based on the defendant's gain rather than the plaintiff's losses. If measured based on the defendant's gain, economic damages are the difference between (1) the defendant's actual economic position as a result of the harmful act, and (2) the defendant's economic position assuming that the alleged harmful act had not occurred.

Figure 1. Damages studies generally adopt this framework. We therefore use it as the basic model for this reference guide.



determining their value as of the valuation date. The plaintiff may be entitled to prejudgment interest through the judgment date on their damages as of the valuation date.³

Isolating the Effect of the Harmful Act

The first step in a damages study is the translation of the legal theory of the harmful event into an analysis of the economic impact of that event. As noted above, in most cases, the analysis considers the difference between what the plaintiff's economic position would have been if the harmful event had not occurred and the plaintiff's actual economic position as of the valuation date.

The characterization of the harmful event begins with a clear statement of what occurred in the actual world. This characterization also will include a description of the but-for world, including the defendant's proper actions in place of their unlawful actions and a statement about the economic situation of all relevant parties absent the wrongdoing, with the defendant's proper actions replacing the unlawful ones. Damages measurement then determines the plaintiff's hypothetical economic position in the but-for world. Economic damages are the difference between that value and the actual value that the plaintiff achieved.

3. The plaintiff may also be due postjudgment interest, which is interest on the damages award from the date of judgment to the date the defendant satisfies the judgment.

Because the but-for world differs from what actually happened only with respect to the harmful act and its reasonably foreseeable economic consequences, damages measured in this way isolate the change in the plaintiff's economic position caused by the harmful act and exclude any change in the plaintiff's economic position arising from other causes.⁴ Thus, a proper construction of the but-for world and measurement of the plaintiff's economic position in the but-for world by definition include in damages only the loss caused by the harmful act.

Properly translating the legal theory of the harmful act into an analysis of the economic impact of that event is a fundamental step in the estimation of economic damages. Failure to do so may result in the damages analysis being discounted or excluded altogether. For example, in *Comcast Corp. v. Behrend*, the U.S. Supreme Court found that the plaintiffs' expert had put forth a damages model that "failed to measure damages resulting from the particular anti-trust injury on which . . . liability in this action is premised."⁵ Citing the discussion in the prior edition of this reference guide, the Court determined that the expert's "methodology . . . identifies damages that are not the result of the wrong" and measures damages "caused by factors unrelated to an accepted theory of . . . harm."⁶

4. Reasonable foreseeability is a legal doctrine that courts use to determine whether the consequences of a legal or contractual breach are too remote to be borne by the defendant. Reasonable foreseeability is not a measure of damages, but rather an upper limit on damages. The use of "reasonable" here relates to the legal concept of the "reasonable person" or "economic person," and concerns actions that a reasonable person or economically rational actor under similar circumstances would take. See, e.g., Richard A. Posner, *Economic Analysis of Law* (2d ed. 1977); Steven Shavell, *Foundations of Economic Analysis of Law* (2004); Israel Gilead & Michael D. Green, *Positive Externalities and the Economics of Proximate Cause*, 74 Wash. & Lee L. Rev. 1517 (2017); John Fabian Witt & Morgan Savage, *Foreseeability Conventions*, 44 Cardozo L. Rev. 1075 (2023); Russ VerSteeg, *Perspectives on Foreseeability in the Law of Contracts and Torts: The Relationship Between "Intervening Causes" and "Impossibility"*, 2011 Mich. St. L. Rev. 1497 (2011). Reasonable foreseeability is not a term of art in the field of economics. While determinations regarding reasonable foreseeability may impact the damages analysis, application of the concept is not within the purview of the damages expert. As such, the damages expert may require assumptions from counsel as to whether a particular outcome is a reasonably foreseeable consequence of the alleged harmful act.

5. *Comcast Corp. v. Behrend*, 569 U.S. 27, 46 (2013).

6. *Id.* at 38 ("The first step in a damages study is the translation of the legal theory of the harmful event into an analysis of the economic impact of that event." The district court and the court of appeals ignored that first step entirely" (citations omitted)). See also *In re POM Wonderful LLC*, No. ML 10-02199 DDP RZX, 2014 WL 1225184, at *5 (C.D. Cal. March 24, 2014) (damages expert "made no attempt, let alone an attempt based upon sound methodology, to explain how Defendant's alleged misrepresentations caused any amount of damages"); *In re Domestic Drywall Antitrust Litig.*, No. 13-MD-2437, 2017 WL 3700999, at *14 (E.D. Pa. Aug. 24, 2017) (expert's damages model was "riddled with assumptions that divorce the model from the facts").

General Categories of Damages Measures

In most cases, damages are measured based on one or more of the following six categories: (1) expectation, (2) reliance, (3) restitution, (4) statutory, (5) liquidated, and (6) punitive.⁷ We discuss each of these in turn.

1. *Expectation*: Plaintiff restored to the same financial position as if the defendant had performed as promised.

Expectation damages typically apply to breach-of-contract claims, where the wrongdoing is the failure to perform as promised and the but-for scenario hypothesizes the absence of that wrongdoing—that is, proper performance by the defendant. Expectation damages are an amount sufficient to place the plaintiff in the same economic position as if the defendant had fulfilled the promise or bargain.⁸

2. *Reliance*: Plaintiff restored to the same position as if the relationship with the defendant or the defendant's misrepresentation (and resulting harm) had not existed in the first place.

Reliance damages generally apply to torts and to some breach-of-contract claims. Such damages restore the plaintiff to the same financial position they would have enjoyed absent the defendant's conduct.⁹ Reliance most often includes out-of-pocket costs,¹⁰ but may also include compensation for lost opportunities when appropriate.¹¹ In such cases, reliance damages may approach expectation

7. The law imposes several limitations on damages, which are discussed below in the section titled "Limitations on Damages."

8. See John R. Trentacosta, *Damages in Breach of Contract Cases*, 76 Mich. Bus. J. 1068, 1068 (1997) (describing expectation damages as damages that place the injured party in the same position as if the breaching party completely performed the contract); *Spring Creek Expl. & Prod. v. Hess Bakken Inv.*, 887 F.3d 1003, 1026 (10th Cir. 2018) (defining expectation damages as the amount of damages necessary to place the injured party in the same position it would have occupied had the breach not occurred). See also Restatement (Second) of Contracts § 344(a) (1981).

9. See E. Allan Farnsworth, *Legal Remedies for Breach of Contract*, 70 Colum. L. Rev. 1145, 1148 (1979) (the objective of reliance damages is to put the promisee, or nonbreaching party, back into the position it would have been had the promise not been made). See also Restatement (Second) of Contracts § 344(b) (1981). Reliance damages include expenditures made in preparation for performance and performance itself. Restatement (Second) of Contracts § 349 (1981).

10. Out-of-pocket costs are expenses incurred by a party in preparing to perform (or in performing) a contract in reliance on the terms of the agreement. This is distinct from the notion of out-of-pocket damages used in securities fraud damages, which relates to the difference between price paid and a but-for price that would have been paid. See the discussion in the section titled "Estimation of Economic Damages in Securities Cases" below.

11. Compensation for lost opportunities embodies the economic concept of opportunity cost, which expresses the relationship between scarcity and choice. Given an economic actor facing a

damages.¹² For a tort, the goal of reliance damages is to place the plaintiff in a position economically equivalent to their position absent the harmful act.¹³ For a breach of contract, measuring damages as the amount of compensation needed to return the plaintiff to the position they would have been in had the contract had not been made in the first place will result in refunding the part of the plaintiff's reliance investment that cannot be recovered in other ways. Thus, reliance damages may be appropriate where the plaintiff made an investment relying on the defendant's performance.

Example: Agent contracts with Owner for Agent to sell Owner's farm. The asking price is \$1,000,000, and the agreed fee is 6%. Agent incurs costs of \$1,000 in listing the property. A potential buyer offers the asking price, but Owner withdraws the listing. Agent calculates (expectation) damages as \$60,000, the agreed fee for selling the property. Owner calculates (reliance) damages as \$1,000, the amount that Agent spent to advertise the property.

3. *Restitution:* Plaintiff compensated by the amount of the defendant's gain from the unlawful conduct, also called compensation for unjust enrichment, disgorgement of ill-gotten gains, or compensation for unbar-gained-for benefits.

Restitution damages are awarded to prevent the defendant's unjust enrichment from the unlawful conduct. The goal of restitution damages is to restore the defendant to the economic position they would have been in had they not engaged in the wrongful act, and typically includes disgorgement of defendant's

choice between mutually exclusive alternatives, the opportunity cost of the alternative selected (e.g., entering into a contract with vendor A) is the value of the best alternative *not* chosen (e.g., entering into a contract with vendor B, C, or D). *See, e.g.,* N. Gregory Mankiw, *Principles of Microeconomics* 6 (8th ed. 2016) ("The opportunity cost of an item is what you give up to get that item." (emphasis omitted)).

12. In some situations, reliance damages can exceed expectation damages, but the court may limit damages to the expectation measure. *See, e.g.,* *Fairholme Funds, Inc. v. Fed. Hous. Fin. Agency*, No. 1:13-cv-1053-RCL, 2022 WL 11110584, at *4 (D.D.C. Oct. 19, 2022) (Plaintiff claimed reliance damages "many multiples higher" than the allowed measure of expectation damages. The court stated, "In contract cases, expectation damages are preferred, with reliance damages considered as an alternative or proxy if expectation damages are not readily ascertainable. . . . [P]laintiffs . . . cannot recover reliance damages so far in excess of ascertainable expectation damages that they would necessarily place those plaintiffs in a better position than they would have been in had the contract been performed.").

13. *See, e.g.,* *East River Steamship Corp. v. Transamerica Delaval Inc.*, 476 U.S. 858, 873 n.9 (1986) ("tort damages generally compensate the plaintiff for loss and return him to the position he occupied before the injury").

ill-gotten gains.¹⁴ In practice, restitution is often the same, from the perspective of quantification, as reliance damages. If the only loss to the plaintiff from the defendant's harmful act arises from an expenditure that the plaintiff made that cannot otherwise be recovered, the plaintiff receives compensation equal to the amount of that expenditure.¹⁵

4. *Statutory*: Plaintiff's compensation is an amount established by statute, either as a set amount per occurrence of wrongdoing or determined using a statutory formula. For example, this measure of damages is often found in cases involving violations of state labor codes and in copyright infringement.
5. *Liquidated*: Plaintiff compensated by an amount specified in a legal agreement or contract between the parties.
6. *Punitive*: Compensation to reward the plaintiff for detecting and prosecuting the wrongdoing or to deter similar wrongdoing in the future.

A plaintiff cannot normally seek punitive damages in a claim for breach of contract but may seek them in addition to compensatory damages in connection with a tort claim.¹⁶ Although punitive damages are rarely the subject of expert testimony, economists have advanced the concept that punitive damages compensate a plaintiff who brings a case for a wrongdoing that is hard to detect or hard to prosecute.¹⁷

Basic Issues Arising in Damages Estimation

Losses Caused

The analysis of losses caused is typically a key part of the analysis of damages. Analyzing damages requires isolating the portion, if any, of the plaintiff's losses that can be attributed to the harmful act from the portion that can be attributed to other factors.

14. Note the distinction between restitution and reliance damages. The objective of restitution damages is to put the promisor, or breaching party, back in the position in which it would have been had the promise not been made. In contrast, reliance damages seek to put the promisee, or nonbreaching party, back in the position in which it would have been if the promise had not been made. See E. Allan Farnsworth, *Legal Remedies for Breach of Contract*, 70 Colum. L. Rev. 1145, 1148 (1979). Both measures seek to restore the status quo ante. See also Restatement (Third) of Restitution and Unjust Enrichment (2011).

15. See Restatement (Second) of Contracts § 344(c) (1981).

16. Compensatory damages are damages "sufficient in amount to indemnify [or compensate] the injured person for the loss suffered." *Damages*, Black's Law Dictionary (11th ed. 2019).

17. See Steven Shavell, *Foundations of Economic Analysis of Law* 244 (2004).

Example: Worker who smokes is the victim of a disease caused either by exposure to xerxium or by smoking. Worker makes leather jackets tanned with xerxium. The disease precludes Worker from earning wages. Worker sues the producer of the xerxium, Xerxium Mine, and calculates damages as all lost wages. Defendant Xerxium Mine, in contrast, attributes most of the losses to smoking and calculates damages as only a fraction of lost wages.

Frequently, the defendant will calculate damages on the premise that the harmful act had no causal relationship to the plaintiff's losses—that is, the plaintiff's losses would have occurred irrespective of the harmful act. The defendant's but-for scenario will thus describe a situation in which the losses happen anyway. This is equivalent to arguing that the harmful act occurred but the plaintiff suffered no incremental losses from it.

Example: Contractors conspired to rig bids in a construction deal. State seeks damages for subsequent higher prices. Contractors' damages estimate is zero because they assert that the only effect of the bid rigging was to determine the winner of the contract and that prices were not affected.

In other situations, unexpected events occurring after the harmful event (known as subsequent unexpected events) can increase or decrease the plaintiff's actual loss relative to what might have been expected at the time of the harmful event.

Example: Housepainter uses faulty paint, which begins to peel a month after the paint job. Owner measures damages as the cost of repainting. Painter disputes this measure on the ground that a hurricane that occurred three months after the paint job would have ruined a proper paint job anyway.

A plaintiff may argue that a harmful act has caused significant losses for many years. The defendant may respond that most of the losses that occurred from the injury are the result of causes other than the harmful act. Thus, the defendant may argue that the injury was caused by multiple factors, only one of which was the result of the harmful act, or that the plaintiff's injury was caused by subsequent events.

Example: Real Estate Agent is wrongfully denied affiliation with Broker. Agent's damages study projects past earnings into the future at the rate of growth of the previous three years. Broker's study projects that earnings would have declined even without the breach because of a downturn in the real estate market.

Comment: The difference between a damages study based on extrapolation from the past, here used by Agent, and a study based on actual data after the harmful act, here used by Broker, is one of the most common sources of disagreement in damages. This is a factual dispute that hinges on Broker demonstrating that there is a relationship between real estate market conditions and the earnings of agents. The example also illustrates how subsequent unexpected events can affect damages calculations.

Hypothesizing Legitimate Conduct of the Defendant in Analyzing the Plaintiff's Economic Position But For the Harmful Event

One party's damages analysis may hypothesize the absence of any act of the defendant that influenced the plaintiff, whereas the other party's damages analysis may hypothesize an alternative, legal act. This type of disagreement is particularly common in antitrust and intellectual property disputes. Although disagreement over the alternative scenario in a damages study is generally a legal question, opposing experts may have been given different legal guidance and therefore made different economic assumptions, resulting in major differences in their damages estimates.

Example: Defendant Copier Service's long-term contracts with customers are found to be unlawful because they create a barrier to entry that maintains Copier Service's monopoly power. Plaintiff Rival's damages study hypothesizes no contracts between Copier Service and its customers, so Rival would face no contractual barrier to bidding those customers away from Copier Service. Copier Service's damages study assumes medium-term contracts with its customers and argues that these would not have been found to be unlawful. Under Copier Service's assumption, Rival would have been much less successful in bidding away Copier Service's customers, and damages are correspondingly lower.

Comment: Assessment of damages will depend greatly on the substantive law governing the injury. The proper characterization of Copier Service's alternative conduct usually is an economic issue. However, the expert must also have legal guidance as to the proper legal framework for damages. Counsel for the plaintiff may instruct the plaintiff's damages expert to use a legal framework different from that used by counsel for the defendant.

Considering All the Differences in the Plaintiff's Situation in the But-For Scenario Versus Assuming That Many Aspects Would Be the Same as in Actuality

The analysis of some types of harmful events requires consideration of effects that involve changes in the economic environment caused by the harmful event, such as price erosion.¹⁸ For a business, the main elements of the economic environment that may be affected by the harmful event include the prices charged by rivals, the demand facing the seller, and the prices of inputs. For example, misappropriation of intellectual property may cause lower prices because products produced with the misappropriated intellectual property compete with products sold by the owner of the intellectual property.¹⁹

In contrast, some types of harmful events do not change the plaintiff's economic environment. The theft of a small amount of the plaintiff's products would not change the market price of those products, nor would an injury to a worker change the general level of wages in the labor market. A damages study need not analyze changes in broader markets when the harmful act plainly has minuscule effects in those markets.

In situations where the plaintiff claims price erosion, the plaintiff may assert that absent the defendant's wrongdoing, a higher price could have been charged and therefore that the defendant's harmful act has eroded the market price. The defendant may reply that the higher price would lower the quantity sold. The parties may then dispute how much the quantity would fall as a result of higher prices.

Example: Valve Maker infringes Rival's patent. Rival calculates lost profits as the profits Rival would have made, including a price-erosion effect. The amount of price erosion is the difference between the higher price that Rival would have been able to charge absent

18. See, e.g., *Beijing Choice Elec. Tech. Co. v. Contec Med. Sys. USA Inc.*, No. 18 C 0825, 2020 WL 1701861, at *9 (N.D. Ill. Apr. 8, 2020) ("To prove lost profits damages, a patentee must reconstruct the market that 'would have developed absent the infringing product,' and the alleged infringer's market share may be a relevant data point in this reconstruction. For instance, if an alleged infringer's market share increases after it introduces the product accused of infringement, that may shed light on what the patentee's market share would have been had there been no infringement. Or, if an alleged infringer's market share does not change significantly after it introduces the accused product, that may indicate that lost profits damages are not appropriate because consumers do not care whether the product uses the patented features at issue" (citations omitted)).

19. See, e.g., *Power Integrations, Inc. v. Fairchild Semiconductor Int'l, Inc.*, 711 F.3d 1348, 1382–83 (Fed. Cir. 2013) ("Lost revenue caused by a reduction in the market price of a patented good due to infringement is a legitimate element of compensatory damages. Indeed, an infringer's activities do more than divert sales to the infringer. They also depress the price [of the patented product].").

Reference Guide on Estimation of Economic Damages

Valve Maker's presence in the market and the actual price. The price-erosion effect is that price difference multiplied by the combined sales volume of Valve Maker and Rival. Defendant Valve Maker counters that the volume would have been lower had the price been higher, and measures damages using the lower volume.

In more complicated situations, the damages analysis may need to focus on how an entire industry would be affected by the absence of the defendant's wrongdoing. For example, one federal appeals court held that a damages analysis for exclusionary conduct must consider that absent the defendant's conduct, firms other than the plaintiff may have entered the market. Competition from such firms would have reduced prices, and as a result the plaintiff's profits would have been lower than those posited in the plaintiff's damages analysis.²⁰

Example: Printer Maker has used unlawful means to exclude rival suppliers of computer printer ink cartridges. Rival calculates damages on the assumption that they would have been the only additional seller in the market absent the exclusionary conduct and that Rival would have been able to sell its cartridges at the same price actually charged by Printer Maker. Printer Maker counters that other sellers would have entered the market and driven the price down, and so Rival has overstated their damages.

Comment: Increased competition lowers prices in all but the most unusual situations. Again, determination of the number of entrants that may have been attracted to the market absent the exclusionary conduct, and analysis of their effect on prices, requires a full economic analysis of the industry.

A clear statement of the plaintiff's situation but for the harmful event is helpful in avoiding double counting that can arise if a damages study confuses or combines reliance and expectation damages.

Example: Marketer is the victim of defective products made by Manufacturer; Marketer's business fails as a result. Marketer's damages

20. See *Cave Consulting Grp., Inc. v. OptumInsight, Inc.*, No. 15-cv-03424-JCS, 2020 WL 127612, at *7 (N.D. Cal. Jan. 10, 2020) (no evidence provided that would provide the jury with any basis to disentangle "favorable aspects of an anticompetitive market such as an unnatural price differential between [plaintiff] and [its] competitors and limited competition from third parties because of the difficulty of entering the market."). See also *Sun Microsystems Inc. v. Hynix Semiconductor Inc.*, 608 F. Supp. 2d 1166 (N.D. Cal. 2009) (holding that a before-and-after damages analysis requires "some showing that the market conditions in the two periods were similar but for the impact of the violation"); *Dolphin Tours, Inc. v. Pacifico Creative Serv., Inc.*, 773 F.2d 1506, 1512 (9th Cir. 1985).

study adds together the out-of-pocket costs of creating the business and the projected profits of the business had there been no defects. Manufacturer's damages study measures the difference between the profits Marketer would have made absent the defects and the profits Marketer actually made.

Comment: Marketer has mistakenly added together damages from the reliance and expectation measures of damages. Under the reliance measure, Marketer is entitled to be put back to where they would have been had they not started the business in the first place. Damages would be total outlays less the revenues actually received. Under the expectation measure, Marketer is entitled to the extra profits they would have received had there been no product defects. Out-of-pocket expenses of starting the business would have no effect on expectation damages because they would be present in both the actual and but-for worlds and thus would offset each other.

Valuation Issues in the Analysis of Economic Damages

As discussed in the previous section, the analysis of economic damages compares a plaintiff's economic position in the but-for world—that is, absent any past and future effects of the alleged harmful act on the plaintiff's economic position—to that plaintiff's economic position in the actual world, and quantifies the economic value of that difference as of a specific date (the valuation date).²¹ Because an analysis of economic damages is, by definition, a valuation exercise, it typically requires that economic experts resolve issues arising from the application of valuation approaches to specific case circumstances. This section introduces valuation issues often addressed by economic experts.²²

21. As noted in the section titled “The Standard General Approach to Quantification of Economic Damages” above, in some situations the damages analysis aims to quantify the defendant's gain instead of the plaintiff's loss. In those situations, the relevant comparison is between the defendant's economic position in the actual world and the defendant's economic position in the but-for world—i.e., absent any past and future effects of the alleged harmful act on the defendant's economic position. The economic issues raised in this section similarly apply when analyzing the defendant's gain from the alleged harmful act.

22. In the section titled “Damages Issues in Selected Types of Cases” below, we introduce the analysis of economic damages in various case types, namely, consumer product liability, securities, antitrust, and loss of personal income. There, we do not specifically address business valuation disputes. However, the methodologies discussed in this section are also relevant to the analysis of economic damages in the context of business valuation disputes.

Issues Arising When the Potential Impact of the Alleged Harmful Act Is Analyzed Directly on a Particular Date

Issues Arising in Connection with the Use of the Market Price of the Product or Asset at Issue

Because a market price is simultaneously a price that the buying party was willing to pay at the time of the transaction and a price that the selling party was then willing to receive in exchange for the asset or product at issue, a market price provides an objective measure of economic value as of a particular point in time. However, as discussed below, the economic value represented by a market price reflects the specific circumstances under which the transaction occurred. Thus, there is often disagreement among experts as to whether and how the market price of the product or asset at issue can be used to analyze damages. This section summarizes some of the key issues that can generate disagreement.

The market price of the product or asset at issue in litigation is a frequent input to the analysis of damages on a particular date. A simple example of such a situation would be where the defendant's negligence caused total destruction of the plaintiff's cargo of wheat, worth \$17 million at the then-current market price of wheat. In this simple example, a damages expert may calculate damages as just that amount, \$17 million.

Market prices are often used to analyze the plaintiff's economic position in the but-for world, the actual world, or both. For example,

- In the simple example above, involving a destroyed cargo of wheat, the market price of wheat at the time of the defendant's negligence may be used to value the plaintiff's economic position but for the alleged harmful act (i.e., the \$17 million value of the plaintiff's cargo of wheat had it not been destroyed).
- In a consumer fraud case involving a product falsely represented to have certain desirable characteristics, the market price of the product may be used to measure the plaintiff's actual economic position at the time of purchase (i.e., how much the plaintiff paid for the product).
- In a defamation case involving a publicly traded company plaintiff, the change in the company's stock price around the time of the defendant's defamatory act may serve as an initial input to the analysis of the difference in the plaintiff's economic position between the but-for world (e.g., using the stock price shortly before the defamatory act) and the actual world (e.g., using the stock price shortly after the defamatory act).

However, disagreement often arises among experts as to whether market prices (or changes in market prices over certain periods) serve as an adequate input to analyzing the impact of the alleged harmful act on the plaintiff's economic position.

First, when using market prices to value the plaintiff's economic position in the but-for world, disagreements may arise as to whether (and, if so, by how much) the market price that the expert relied on for their analysis represents a value that was itself impacted by the alleged harmful act. If the value represented by the market price reflects the impact of the alleged harmful act, such impact must be removed from the market price in order to properly value the plaintiff's economic position in the but-for world.

For example, if the destroyed cargo of wheat mentioned above was large enough that its destruction significantly constrained the supply of wheat, thereby increasing the market price of wheat, then the market price of wheat immediately following the destruction of the cargo of wheat will overstate the plaintiff's economic position but for the alleged harmful act and, thus, will overstate damages. In that case, disagreement may arise among the experts as to the extent to which the market price of wheat was affected by the defendant's negligence. Moreover, disagreement may arise with respect to what statistical or economic methodologies should be used to quantify the necessary adjustment to the market price in order to account for that impact.

Second, when market prices are used to analyze the plaintiff's economic position in the actual world, disputes may arise as to whether (and, if so, by how much) the market price fully reflects the impact of the alleged harmful act. If the market price does not fully reflect the impact of the alleged harmful act, then it needs to be adjusted using a reliable methodology to properly value the plaintiff's economic position in the actual world.

For example, suppose that a manufacturer of wood windows with publicly traded common stock treats its windows with a defective preservative, causing the windows to rot. The window manufacturer sues the preservative manufacturer for damages from lost sales and from the cost of replacing the defective windows. The window manufacturer's expert may be tempted to use the window manufacturer's stock price following the disclosure of the alleged harmful act as a measure of the manufacturer's economic position in the actual world. A potential problem with using the market price of the plaintiff's stock after the alleged harmful act is that the stock market may anticipate recovery in the form of a damages award, and this will attenuate at least some of the decline in price following the alleged harmful act.²³ In the extreme, if stock traders expect that

23. Note that this understatement of damages arises when the publicly traded company stands to recover a damages award as the plaintiff in a matter. However, changes in the market price of company stock have a different role in situations, such as a securities matter, where the public company is the defendant. Damages may be overstated by an unknown amount if the release of

the plaintiff will receive exactly full compensation, the plaintiff's market value is unlikely to change at all when knowledge of the wrongdoing—including the fact that a damages award will be made—hits the stock market. Thus, in this example, the use of the observed market price of the plaintiff company's stock at the time of the injury understates the actual amount of harm by an unknown amount, so the expert should consider using additional valuation techniques. Disputes may arise among the experts as to which valuation techniques should be used to quantify the harm and how they should be applied.

Third, when market prices are used to analyze the plaintiff's economic position, disputes may arise as to whether (and, if so, by how much) market prices reflect not just the impact of the alleged harmful act, but also the impact of other factors. Any impact of factors other than the alleged harmful act must be removed from the estimate of economic damages in order to isolate the harm caused by the alleged harmful act.

For example, consider a securities fraud matter where a mining company defendant is alleged to have overstated the size of its mineral reserves and where the company's stock price declined by 30% following its revelation of the true size of the mineral reserves. Suppose that, on the day the company revealed the true size of its reserves, it also announced that a natural disaster had caused its largest mine to collapse. To quantify the losses, if any, caused by the market learning of the company's true mineral reserves, it is necessary to disaggregate the impact of that disclosure from the impact of the disclosure of the mine collapse. A dispute may arise as to how the change in stock price should be adjusted to remove the impact of the disclosure of the mine collapse and isolate the impact of the disclosure of the company's true reserves.

Issues Arising in Connection with the Use of the Market Price of Ostensibly Similar Products or Assets

In many cases, experts cannot take the market price of the product or asset at issue and apply it directly to analyze the plaintiff's economic position in the but-for world or in the actual world. For example, there may not be a market for the product or asset at issue from which to obtain a market price (e.g., a company may not be publicly traded). However, experts may still be able to use the market prices of ostensibly similar products or assets—often labeled comparables—to analyze the economic value of the product or asset at issue.

previously fraudulently concealed adverse information causes a reduction in the value of the company both because of the adverse information and because the market anticipates that the company will pay a large damages award to investors who overpaid for their shares during the period when the information was concealed.

Identifying products or assets that are ostensibly similar to the allegedly affected product or asset

A common area of disagreement when relying on the market price of comparables to analyze damages is whether the identified comparables are economically similar to the allegedly affected product or asset. For example, in matters involving real estate, experts often identify comparables from the universe of nearby properties with similar characteristics (e.g., property size, condition, proximity to amenities, zoning restrictions). If the identified comparables are not similar to the product or asset at issue along economically relevant dimensions, the comparables' market prices may not serve as a reliable measure of the economic value of the allegedly affected product or asset (in the but-for world or in the actual world) and may lead the economic expert to significantly misestimate economic damages.

There may also be a dispute as to whether the period when the identified comparables were traded (and thus when their market price was determined) is too distant from the damages valuation date. Temporal distance between the damages valuation date and the period when the selected comparables were traded may cause the market prices of the selected comparables to reflect macroeconomic and business conditions that are significantly different from those in place as of the damages valuation date. In that case, the market prices of the comparable products or assets may not reliably measure the economic value of the product or asset at issue. For example, in a matter involving a sale of a private oil company that took place at a time when oil prices were at historic lows, evaluating the fairness of the sale price by comparing it to the prices at which similar companies were sold during an earlier period (when oil prices were significantly higher) could lead the expert to overstate the value of the company as of the sale date.

Adjusting the market price of ostensibly similar products or assets

Even when experts agree on a list of comparables to use in the damages analysis, disputes may arise as to whether (and, if so, how) the value of the allegedly affected product or asset implied by the comparables' market prices needs to be further adjusted to adequately measure the economic value of the product or asset at issue.

Accounting for differences between the ostensibly similar products or assets and the allegedly affected product or asset. Despite experts' best efforts, circumstances are often such that the best available comparables remain different from the product or asset at issue in economically significant ways. In such circumstances, additional adjustments to account for those differences may

be required to ensure that the economic value of the allegedly affected product or asset is not misestimated. For example, in a matter involving a business valuation, the business at issue may be much smaller than any of the identified comparable businesses, rendering value comparisons inappropriate. Therefore, an expert may propose to restate the value of comparable businesses using valuation ratios—a ratio of some measure of value, such as stock price, to some measure of economic output, such as earnings—and multiply the comparables' valuation ratios by the measure of economic output for the smaller business to estimate its value, thereby adjusting for size.

However, disputes often arise over whether any adjustments made to the value of the comparables appropriately capture all significant economic differences between the comparables and the product or asset at issue. In the business valuation example, a dispute may arise as to whether the adjustment based on valuation ratios fully accounts for differences between the company at issue and the comparable companies, such as differences in profitability, growth prospects, or riskiness.

Example: Oil Company deprives Gas Station Operator of the benefits of Operator's business. Oil Company's damages study starts by calculating the ratio of gas station sale value to gasoline sales for five nearby gas station businesses that have sold recently. The average ratio is \$0.26 per gallon of sales per year. The Operator sells 1.6 million gallons per year, so the business was worth $\$0.26 \times 1,600,000 = \$416,000$, according to the Oil Company's expert. The expert for the Operator argues that the sales used by the Oil Company occurred before a major business relocated nearby. Thus, the gas station sale value to gasoline sales should be increased to \$0.30 to reflect the new growth rate as a result of the expected increase in business. The Operator's expert calculates the business to be worth $\$0.30 \times 1,600,000 = \$480,000$.

Accounting for the impact of the harmful act. In some cases where experts rely on comparables to analyze damages, the alleged harmful act did not cause a total loss of the value of the product or asset at issue, but rather caused only a reduction in its value. In those cases, damages experts may adapt the valuation implied by the comparables to measure the loss caused by the alleged harmful act—that is, the difference between the plaintiff's economic position in the but-for world and in the actual world. For example, in a business valuation case, an expert may propose to use valuation ratios to analyze damages, as illustrated by the numerical example below, although disputes may arise as to their appropriateness.

Example: Oil Company breaches an earlier agreement with Gas Station Operator and opens another station near Operator's station. As a result, Operator's gasoline sales are reduced by 700,000 gallons per year. Oil Company's damages study applies an average valuation ratio (based on recent sales of gas station businesses) of \$0.26 per gallon of sales per year to the reduction in sales: $\$0.26 \times 700,000 = \$182,000$. In contrast, Operator's damages study uses a statistical analysis (also based on recent sales of gas station businesses) that finds the impact of sales on gas station value depends on the total sales volume of the gas station, such that, given the Operator's level of sales absent the breach, each additional (or marginal) gallon of *subtracted* sales reduces value by \$0.47. As a result, the Operator's expert calculates damages of $\$0.47 \times 700,000 = \$329,000$.

Comment: Because fixed costs associated with running a gas station (e.g., rent) are diluted by larger sales amounts, the average valuation of gasoline sales will be less than the marginal valuation. Thus, a damages model that accounts for the level of sales absent the breach is the conceptually correct approach.

Issues Arising in Connection with the Use of Models to Simulate the But-For Price

In certain cases, experts cannot adjust the market price of the asset or product at issue—or of ostensibly similar assets or products—to reflect the impact of the alleged harmful act. In these cases, some experts have used statistical and economic models to simulate market prices in the but-for world.

It is well established in economics that market prices are determined by supply and demand. To estimate market prices in the but-for world, experts have proposed models that seek to represent the price-formation process, or the outcome of that process, in the relevant market when supply and demand for the asset or product at issue exclude the effects of the alleged wrongdoing. Experts rely on assumptions that simplify some of the complexities of real markets and focus their analysis on the specific impact of the harmful act.

Experts also rely on a wide variety of data to estimate the relevant variables and parameters of their models and represent real markets to a reasonable degree of accuracy. These data can include market data such as prices and quantities sold of the asset or product at issue, or the prices and quantities sold of related assets or products (such as substitute and complementary products). Experts can also rely on nonmarket data, such as accounting data or data generated via surveys.

Disagreements occur with respect to the assumptions and data used in these models. This section summarizes certain issues that arise when experts rely on statistical and economic models to simulate or approximate but-for prices.

Realistic modeling of preferences, incentives, and constraints of buyers and sellers

Experts use economic and statistical models to simulate real markets or market outcomes by focusing on the analysis of certain key elements and relationships of the market and by making assumptions about the influence of other less relevant characteristics. For a model to be useful, it needs to provide a sufficiently realistic representation of the relevant market. This usually involves defining the key characteristics of supply and demand of the asset or product at issue. Disputes may arise over which assumptions about demand and supply more realistically reflect the relevant market in the but-for world.

A valid model must provide a sufficiently realistic representation of the preferences of buyers. Experts usually achieve this by relying on methods of demand estimation commonly used by economists and marketing academics. These methods are broadly classified in two categories. The first category, known as the revealed preference method, relies on data from the choices made and prices paid by buyers in the real world to infer the underlying preferences and restrictions of buyers. The second category, known as the stated preference method, relies on data from surveys of relevant subjects that are considered representative of the population of buyers of the asset or product at issue.

When data from surveys are used, disputes may arise about whether such data accurately represent respondents' preferences—respondents may behave differently when taking a survey compared to how they would behave in the real market. Disputes may arise about how realistically a survey represents the environment in which buyers make purchase decisions in the real world, or about how accurately the sample of survey respondents represents the relevant population of buyers. Refer to the *Reference Guide on Survey Research*, in this manual, for a detailed treatment of these and other issues that arise when surveys are conducted.

A common question that arises when experts rely on a model to represent demand for an asset or product is whether the model appropriately accounts for variation in the preferences of different groups or types of buyers. For example, in an intellectual property dispute, an expert could use an economic model to estimate the price of a brand-name drug covered by a patent in a but-for world where the patent was not enforceable and generic drug manufacturers were allowed to compete. An overly simplistic model of demand could lead to a prediction that the entry of generic drug manufacturers into the market would result in a decrease in the price of the brand-name drug. However, a more realistic demand model

that takes into account the preferences of different types of consumers could lead to a finding that the market price of the brand-name drug would increase as the smaller segment of consumers who are brand-conscious would be willing to pay more for the brand-name drug.

A valid model must also provide a sufficiently realistic representation of the preferences and incentives of sellers. In many cases, experts rely on simplifying assumptions to represent the characteristics of supply. Some experts have assumed that the supply quantity is fixed at the quantity that was sold in the actual world. Other experts rely on economic frameworks to account for the changes in the economic position of sellers. These frameworks may include simplifying assumptions about the general relationship between the quantity offered and the price offered in the market or may include sophisticated models that simulate the behavior of individual sellers. Disputes commonly arise about whether the assumptions made by an expert regarding supply reflect the true positions and incentives of the suppliers in the but-for world.

Defining a valid model of supply generally requires consideration of key characteristics of the market that may influence the prices that consumers pay. For example, in a dispute involving the alleged mislabeling of a prescription drug's potential side effects, an expert must consider that the prices that consumers pay for prescription drugs may be influenced by insurance companies that can negotiate prices with the manufacturer and that can adjust consumer prices based on copays and deductibles.

It may also be necessary to define the characteristics of key competitors, distributors, retailers, and other relevant market participants. For example, a model that attempts to measure the impact on ticket prices of an alleged anti-competitive agreement between two bus lines should take into account available alternative means of transportation, such as cars, trains, or possibly airplanes. Disputes may arise as to the number, type, and characteristics of relevant alternatives consumers may have.

Modeling the impact of the alleged misconduct on demand and supply

To be helpful in the estimation of damages, a model of but-for prices must adequately reflect supply and demand absent the alleged misconduct while excluding the influence of unrelated factors. In general, the more complex the allegations are, or the more complex the facts surrounding the allegations are, the more difficult it will be for an economic model to reliably estimate market conditions absent the alleged misconduct.

Complex allegations may require more flexible models to represent the but-for world adequately. In general, models that allow more flexibility and that

provide the expert with more control over how to represent demand and supply in the but-for world are more difficult to estimate, and require larger amounts of data. For example, consider a dispute between a large buyer of insurance and an insurance company in which the allegations involve whether the insurer properly invoked a complex set of conditions to cancel a portfolio of policies and avoid paying for them. The damages model would require significant flexibility to perform the valuation of the portfolio in a but-for world where the defendant insurer could invoke some, but not all, of the cancellation conditions that it did in the actual world. Disputes commonly arise about whether an expert model adequately reflects complex allegations.

A model used by an expert should be able to account for the impact of confounding factors unrelated to the allegations but that may result in the same type of harm alleged. For example, in a dispute about potential contamination of groundwater from an industrial plant, an expert could estimate a statistical model—known as a hedonic model—to predict home prices based on certain characteristics of a home (e.g., number of bedrooms, number of bathrooms, etc.). The expert could attempt to use sale price data of homes before and after the alleged contamination event to estimate the impact of the alleged contamination on home values. However, if a zoning change affecting the properties at issue was enacted around the same time as the alleged contamination event, the expert's model would need to distinguish any decrease in the price of homes from the alleged contamination from any decrease (or increase) due to the zoning change. Disputes commonly arise about whether an expert model adequately controls for the impact of confounding factors.

Surveys provide the expert with some control over how to represent the but-for world via the information conveyed to survey respondents. However, disputes may arise about the specific information that respondents are assumed to have in the but-for world. For example, an expert could propose a survey to analyze how consumers would react to a label disclosing the use of genetically modified corn on a package of cereal labeled “100% natural corn.” Experts may disagree about the wording, size, and placement of such a label, all of which can affect the damages estimate arising from the survey.

Accounting for Market Frictions

When relying on market data to analyze economic damages as of a specific date, it may be necessary to consider how market frictions may have impacted market prices—whether it be the price of the allegedly affected product or asset or the price of ostensibly similar products or assets.

Perfectly competitive markets have what economists term a frictionless market structure. These markets have (1) a large number of buyers and sellers

of a single, homogeneous product; (2) fully informed participants; and (3) the feature that participants can easily enter or exit from the market. A friction is anything that prevents the market from being perfectly competitive. Economists have identified various types of market frictions that can impact market outcomes, including market power (e.g., oligopolies), adverse selection, and taxes.²⁴

For example, markets for businesses and properties have frictions that may make transaction values depart from the usual concept of the price negotiated by a willing seller and a willing buyer. In the case of a forced sale, and thus a less willing seller, the transaction price may understate the value. Adverse selection, which occurs when one party knows more about a property or business than the other, may cause severe failures in some markets.²⁵ Sales prices tend to be lower when equipment with hidden defects cannot be distinguished from equipment in unusually good condition. This is because buyers are only willing to pay a lower price that reflects the possibility of hidden defects in the equipment and because owners of equipment in good condition may choose not to offer it at that lower price.

Example: Negligence of Tire Maker causes the total loss of a Boeing 737 airplane. Tire Maker's damages analysis uses the prices of 737s of similar age in the used airplane market to set a value of \$23 million on the ruined airplane. However, Airline offers the testimony of an economist expert who explains that only a small fraction of 737s are ever put up for sale in the used airplane market. Rather, airlines choose to sell only defective airplanes because they continue to fly nondefective 737s. The expert then adjusts for the adverse selection of inferior airplanes in the used airplane market and places a value of \$42 million on the airplane.

Comment: Airline expert's adjustment is merited in principle, although it is challenging to carry out and is likely to be the subject of expert disagreement.

In financial markets, differences in information available to different market participants or limited asset liquidity can lead to market prices that differ, sometimes significantly so, from the prices that would have prevailed in perfectly competitive financial markets. For example, the sale of a large block of shares of

24. A comprehensive treatment of economic frictions is outside the scope of this reference guide.

25. See, e.g., George A. Akerlof, *The Market for "Lemons": Quality Uncertainty and the Market Mechanism*, 84 Q. J. Econ. 488 (1970), <https://doi.org/10.2307/1879431>.

a company's common stock may happen at a price that is much lower than recently prevailing market prices because potential purchasers of that block of shares are concerned that the seller might have negative, material nonpublic information about the value of the shares.

A major source of friction in property and business markets is the capital gains tax. Because capital gains are taxed only at realization, subtracting the amount of unrealized capital gains taxes from the value of a business or property would generally understate the value of the business or property to the existing owners if they have no plans to sell, except in the very distant future.

Issues Arising in Connection with the Use of Repair or Abatement Costs

In some cases, economic damages can be calculated as the costs that the plaintiffs incurred or would incur to restore their economic position from the actual world to the but-for world in which the alleged harmful act did not occur. These costs could include repair costs (e.g., the cost of the parts and labor used to repair a defective product), replacement costs (e.g., the cost of a third party hired to perform a task that the defendant was contracted to do but did not), or the costs of reducing or abating any harm caused by the alleged misconduct (e.g., the cleanup costs following an environmental contamination incident).

Damages experts often disagree on how to calculate these costs. First, experts may disagree on whether any costs allegedly incurred are attributable to the alleged misconduct. For example, if a service hired by the plaintiff to repair an allegedly defective product also performed general maintenance, the total cost paid may overestimate damages. Second, experts may disagree on the costs that are necessary to repair or abate the consequences of the alleged misconduct. For example, in a case where the plaintiff alleges that they will need to install expensive air filtration devices to reduce foul odors from a defective air conditioning system, experts may disagree on whether such costs are necessary if cheaper alternatives (e.g., more frequent maintenance) are available. Third, experts may disagree about whether the potential costs of repair match the economic value of the loss experienced by the plaintiff. For example, consider a builder who in the process of selling a house learned that a contractor used cheaper materials than promised. Assume that the cost of removing the cheaper materials and replacing them with the materials originally specified would be \$100,000; however, the builder was able to sell the home by offering a discount of only \$50,000. In this case, the replacement costs would overstate the economic loss suffered by the builder.

Issues Arising When the Potential Impact of the Alleged Harmful Act Spans Multiple Dates and Damages Are Analyzed Indirectly on a Single Date

Having discussed various issues that arise when experts analyze damages by assessing the impact of the alleged harmful act on a specific date directly, we now turn to issues arising in situations where economic experts analyze damages on a particular date indirectly. Specifically, we turn to situations where experts first assess the direct economic impact of the alleged harmful act (i.e., the flow of economic income or loss at a specific point in time that can be causally linked to the defendant's alleged misconduct) over multiple dates (e.g., the lost profits from operating a facility that was expected to produce widgets for a long period of time but was destroyed by a defendant's misconduct) and then value the cumulative direct impact over multiple dates as of a single valuation date.

Estimating the Period-by-Period Direct Impact of the Alleged Harmful Act

In situations where experts assess the direct economic impact of the alleged harmful act over multiple dates, it is necessary to estimate flows of economic income or loss to the plaintiff in past and future periods that resulted (or were expected to result) from the alleged harmful act. In doing so, experts face a number of considerations.

Many of those considerations, such as how the actions of plaintiffs and defendants would have differed had the alleged harmful act not occurred, or whether all economic consequences of the alleged harmful act have been considered, have been introduced more broadly in the section titled “Valuation Issues in the Analysis of Economic Damages” above. This section further explores these considerations in the specific context of assessing the period-by-period direct economic impact of the alleged harmful act.

Disputes about how the defendant's actions would have differed had the alleged harmful act not occurred

When analyzing damages, experts often rely on the plaintiff's allegations (e.g., as stated in a complaint or statement of claim) as the basis for identifying the alternative actions that the plaintiff alleges the defendant should have undertaken instead of the alleged harmful act. However, in some cases, the plaintiff's

allegations do not specify fully what the defendant could and should have done instead of the alleged harmful act, or the allegations are squarely contradicted by case facts. For example, consider a derivative lawsuit where a company's shareholders sue the CEO for using the firm's funds in wasteful company acquisitions that are expected to reduce the company's quarterly profits for periods lasting well into the future. The plaintiff's allegations may leave unspecified what the CEO should have done with the funds instead of the allegedly wasteful acquisitions or the allegations may assume that certain investment opportunities were available that the record demonstrates were not.

Depending on the circumstances, the ambiguity remaining in the allegations may be significant enough that it causes experts on opposite sides of the litigation to reach substantially different estimates of damages. For example, in the case of the CEO alleged to have engaged in wasteful corporate acquisitions, the expert retained by the CEO defendant may assume that the funds used to finance the acquisitions would have been used to fund internal projects that were expected to be as unprofitable as the wasteful acquisitions. On the other hand, the expert retained by the investor plaintiffs may assume that the funds would have been distributed to the shareholders as quarterly dividends instead. Thus, when assessing opposing experts' damages analyses, a critical issue may be how each expert modeled what the defendant's actions would have been had the alleged harmful act not occurred.

When the plaintiff's allegations are consistent with any of several alternative actions that the defendant might have taken had the alleged harmful act not occurred, there is a risk that expert analyses become tainted by hindsight. With the benefit of hindsight—often, after years have passed since the alleged harmful act occurred—the parties may instruct their experts to assume that the defendant would have taken particular actions (among various possible alternatives that are consistent with the plaintiff's allegations) that lead to a more favorable damages number. However, the defendant's actions, had the alleged harmful act not occurred, would have been based on the defendant's knowledge at the time of the alleged harmful act, and not on whatever knowledge only later became available, including as a result of litigation. In the example above involving wasteful corporate acquisitions, depending on what the CEO contemporaneously knew, had the alleged harmful act not occurred, the CEO's decision could have been either to pay out the funds as dividends (as the shareholders' expert assumed) or to reinvest the funds into internal projects (as the CEO's expert assumed).

Disputes about how the plaintiff's actions would have differed had the alleged harmful act not occurred

Even when a defendant's actions, had the alleged harmful act not occurred, are fully specified, there may be disagreements about what the plaintiff would have done in the hypothetical world where the alleged harmful act did not occur. Such

disagreements may also lead to significantly different damages numbers. Take the numerical example below, where the plaintiff in a real estate matter argued that undeveloped land was taken from him by the defendant's harmful act and that damages should include the value of the undeveloped land, as well as the value of the stream of monthly rental income for years to come that the plaintiff would have earned had the plaintiff been able to develop the land.

Example: Property Owner sues County for the value of undeveloped property condemned for a rapid-transit extension. Property Owner's damages claim is \$18 million, which is the appraisal value of a hypothetical condominium development on the property less the anticipated cost of building the development. The County's expert, an appraiser, argues that the market value of the property is \$2 million, based on comparable undeveloped land nearby.

Comment: In principle, if the real estate market is perfectly competitive, the current market value of undeveloped land and the market value of the same land with proper development, less the cost of that development, should be the same, because buyers would bid for the developed properties based on the value of the undeveloped land. Thus, Property Owner may have understated the development costs. But the value of nearby properties used as comparables may understate the value of the condemned property—they may be available for sale because they lack certain features that would make them more desirable to develop, such as a view. On the other hand, the Property Owner's valuation does not reflect the probability that the Property Owner may not succeed in building the condominium.

Moreover, as was the case with respect to modeling the defendant's actions had the alleged harmful act not occurred, there is a risk that expert analyses may be tainted by hindsight when making assumptions about what actions the plaintiff would have taken, had the alleged harmful act not occurred.

Disputes about whether all economic consequences of the alleged harmful act are being accounted for

Another area where disputes may arise when analyzing the direct economic impact of the alleged harmful act over multiple periods is whether the experts' damages analyses have fully considered all the reasonably foreseeable economic consequences of the alleged harmful act. Failure to fully consider all reasonably foreseeable economic consequences of the alleged harmful act may lead experts to estimate damages incorrectly, possibly significantly so.

Reference Guide on Estimation of Economic Damages

For example, where the injury takes the form of lost sales volume, the plaintiff usually has avoided the cost of production for the lost sales. Thus, an expert would overstate damages by not including the avoided costs of production as a reduction in the estimated direct economic impact of the alleged harmful act. Calculation of these avoided costs is a common area of disagreement about damages. Conceptually, avoided cost is the difference between the cost that would have been incurred at the higher volume of sales but for the alleged harmful act and the cost actually incurred at the lower volume of sales achieved. The avoided-cost calculation is done for each period. The following are some of the issues that arise in calculating avoided cost:

- For a firm operating at capacity, expansion of sales is cheaper in the longer run than in the short run; however, if there is unused capacity, expansion of sales may be cheaper in the short run.
- The costs that can be avoided if sales fall abruptly are less in the short run than in the longer run.
- Avoided costs may include marketing, selling, and administrative costs as well as the cost of manufacturing.
- Some costs are fixed, at least in the shorter run, and are not avoided as a result of the reduced volume of sales caused by the harmful act.

Sometimes putting costs into just two categories is useful: those that vary with sales (variable costs) and those that do not vary with sales (fixed costs). This breakdown is approximate, however, and does not do justice to important aspects of avoided costs. In particular, costs that are fixed in the short run may be variable in the longer run. Disputes frequently arise over whether particular costs are fixed or variable. One side may argue that most costs are fixed and were not avoided by losing sales volume, whereas the other side may argue that many costs are variable.

Certain accounting concepts relate to the calculation of avoided cost. Profit-and-loss statements frequently report the cost of goods sold.²⁶ Costs in this category are frequently, but not uniformly, avoided when sales volume is lower. But costs in other categories, called operating costs or overhead costs, may also be avoided, especially in the longer run. One approach to the measurement of avoided cost is based on an examination of all of a firm's cost categories. The expert determines how much of each category of cost was avoided.

An alternative approach uses regression analysis or some other statistical method to determine how costs vary with sales as a general matter within the firm or across similar firms. The results of such an analysis can sometimes be used to measure the costs avoided by the decline in sales volume caused by the harmful act.

26. See, e.g., *United States v. Arnous*, 122 F.3d 321, 323–24 (6th Cir. 1997) (finding that the district court erred when it relied on government's theory of loss because the theory ignored the cost of goods sold).

Disputes about the use of expected outcomes to estimate the impact of the alleged harmful act

In most situations where the expert's damages analysis includes an evaluation of the impact of the alleged harmful act on future flows of economic income or loss, it is necessary to account for the uncertainty that surrounds the actual amount of those flows under different, yet-to-be-determined economic circumstances. For example, in a matter involving the valuation of a business, it is typically necessary to analyze how the future profits of the business may differ under different states of the world (e.g., in an economic recession versus in an economic boom, or if the business's new product becomes popular versus if it does not).

A fundamental economic principle in the analysis of uncertainty is that the value of future flows of economic income or loss depends on the expected outcome of those flows.²⁷ The expected outcome of an uncertain future flow of economic income or loss is conceptually simple to calculate. First, it is necessary to determine the specific outcome for the economic flow under the various possible scenarios and the probability that each scenario will occur. Then, the expected outcome is calculated as the weighted average of the outcomes for the economic flow across the various possible scenarios, weighting each scenario by the probability that it will occur. For example, in a simple situation where a company faces a one-time future profit of \$1 million with probability 40% and a profit of \$16 million with probability 60%, the company's expected profit is \$10 million (\$1 million times 40% plus \$16 million times 60%).²⁸

Application of the expected-value approach involves a careful study of the various risk factors that are expected to impact the ultimate outcome for the flow of economic income or loss. For example, in a business valuation matter involving a commercial real estate developer, it may be necessary to consider how the company's profitability would differ depending on risk factors that are likely to include future sales prices of commercial real estate, construction costs

27. As discussed in the section titled "Disagreements About the Discount Rate to Value the Impact of the Alleged Harmful Act in Periods After the Valuation Date" below, another important issue in the analysis of damages when there are future flows of economic income or loss is the appropriate discount rate to be used to transform expected future flows into their present value.

28. Sometimes the alternatives and interactions between the possible outcomes become so complex that more sophisticated analytical methods may be required. In such cases, experts often generate hundreds to millions of possible outcomes using probabilistic simulation techniques (e.g., Monte Carlo). These techniques usually start from a model of the relation between the outcome variable of interest (e.g., a company's profits, or the payoffs from a financial derivative) and the risk factors that affect that variable (typically based on findings from the peer-reviewed academic literature or statistical analyses of historical data). Then, the value that the outcome variable of interest would take under different future realizations of the relevant risk factors is simulated. Because those realizations are not known with certainty, they are simulated based on random draws from the assumed or estimated statistical distribution of the risk factors. Those random draws are then used as an input to the model to determine the simulated value of the outcome variable of interest.

(e.g., labor, steel, or lumber), local zoning restrictions, and availability of bank financing. Disputes often arise among experts over which risk factors must be considered in analyzing expected values and with respect to how those factors are likely to impact the flows of economic income or loss under different situations. Economic experts often rely on peer-reviewed academic literature or on statistical analyses based on historical data to determine the relevant risk factors and how they are expected to impact the ultimate outcome.

Disputes about the economic situation in future periods

When it is necessary to estimate the direct period-by-period impact of the alleged harmful act in future periods, disputes often arise over the basis for the assumption that the future will look a certain way. For example, in matters involving the valuation of startup companies with no track record of earlier profitability, a common source of disagreement is the likelihood that the company at issue will reach profitability and, if it does, how profitable it will be. An expert who values a startup company with no track record of earlier profitability under the assumption that the company will reach profitability is likely to be scrutinized on the basis for that assumption.

Example: Plaintiff Xterm is a failed startup. Defendant VenFund has breached a venture-capital financing agreement. Xterm's damages study projects the profits it expected to make under its business plan. VenFund's damages estimate, which is much lower, is based on the value of the startup as revealed by sales of Xterm's equity made just before the breach.

Comment: Both sides confront factual issues in validating their damages estimates. Xterm needs to show that the expected profits according to its business plan were supported by relevant information available to it as of the time of the breach. VenFund needs to show that the sale of equity places a reasonable value on the firm (e.g., it did not take place in a "fire sale" and it incorporated the same information that was available as of the date of breach). This dispute can also be characterized as whether the plaintiff is entitled to expectation damages or reliance damages. Certain jurisdictions may limit damages for firms with no track record of profitability.²⁹

29. See, e.g., *Schonfeld v. Hilliard*, 218 F.3d 164, 172 (2d Cir. 2000) ("Although lost profits need not be proven with 'mathematical precision,' they must be 'capable of measurement based upon known reliable factors without undue speculation.' Therefore, evidence of lost profits from a new business venture receives greater scrutiny because there is no track record upon which to base an estimate. Projections of future profits based upon 'a multitude of assumptions' that require 'speculation and conjecture' and few known factors do not provide the requisite certainty." (citations omitted)).

*Disputes about the consideration of non-cash effects
of the alleged harmful act*

In some situations, a careful assessment of the plaintiff's economic situation in the actual world and in the but-for world reveals that the direct economic impact of the alleged harmful act on a period-by-period basis may include the loss of a flow of economic income that does not directly involve the loss of actual cash.

For example, in some firms, employee stock options are a significant part of total compensation. Employee stock options grant their holders the right to purchase shares of company stock at a fixed price. Stock options are often used by early-stage businesses because the options do not require the business to pay out any cash at the time they are granted. At the same time, the options are valuable compensation for employees because they potentially allow them to benefit from the company's future growth. Specifically, at a future date, the options may be exercised, and the employee option holder will pay only the previously agreed exercise price (typically, the price per share at the time the options are received) as opposed to the price per share at the time the options are exercised. If the company continues to grow successfully into the future, such that the shares' market price vastly exceeds the exercise price, option holders will pay a price for new shares that is a fraction of the shares' market price.

In a lost-sales matter where a plaintiff company's employees are compensated using stock options, the parties may dispute whether the value of options should be included in the costs avoided by the plaintiff as a result of lost-sales volume. The defendant's expert might argue that stock options should be included, because their issuance is costly to the shareholders. The defendant might place a value on newly issued options and amortize this value over the period from issuance to vesting. The plaintiff, in contrast, might exclude options costs because the options cost the firm no cash payout at the time the options are granted. However, compensating employees using stock options imposes real economic costs on the firm's shareholders, in the form of actual value outlays when these options are exercised or in the form of expected future value outlays at the time the options are granted. Those costs should be considered.

Example: Firm A pays its sales manager \$2,000 for every machine sold at \$100,000, as well as options to purchase 1,600 shares in a year at the existing price of \$10 per share. Firm A asserts that it lost \$10 million in sales (100 machines) as a result of Firm B's disparagement of Firm A. In its damages analysis, Plaintiff Firm A states that the lost sales represent lost profits of \$5.8 million: \$10 million less \$4 million in avoided production costs and \$200,000 in avoided sales commissions. Defendant Firm B calculates that each stock option is worth \$5 today based on an analysis

using accepted financial models to value the options. Thus, Firm B asserts that damages are \$5 million: \$10 million less \$4 million in avoided production costs and \$1 million in sales commissions (\$200,000 plus $\$5 \times 100 \times 1,600$).

Valuing the Period-by-Period Impact of the Alleged Harmful Act as of a Single Valuation Date

This section discusses issues that arise when determining the appropriate rates for the purposes of discounting and capitalization, as well as the appropriate valuation date. Once an expert generates estimates of the direct economic impact of the alleged harmful act in each relevant period, it is then necessary to convert those estimates into a single measure of damages as of a specific date—the damages valuation date. As noted above, experts do so by applying the fundamental economic principle that the value of flows of economic income or loss at different points in time can be translated into an equivalent value as of a particular date through discounting or capitalization using appropriate rates.

Discounting is the act of translating a future economic value to its equivalent economic value as of an earlier point in time. For example, discounting at an appropriate rate can be used to translate receiving \$105 a year from now into its economic value today and to compare that to receiving \$100 today. If the appropriate discount rate is 10%, then today's value of receiving \$105 in a year is approximately \$95—calculated as \$105 divided by $(1 + 10\%)$ —and receiving \$100 today is worth more than receiving \$105 in a year. Capitalization is the reverse of discounting—that is, it is the act of translating a past economic value into its equivalent value as of a later point in time. For example, capitalizing at an appropriate rate can be used to compare receiving \$100 today to having received \$95 a year ago (and having reinvested that amount). If the appropriate capitalization rate is 10%, then today's value of \$95 received a year ago is approximately \$105—calculated as \$95 multiplied by $(1 + 10\%)$ —and \$95 received a year ago is worth more than \$100 received today.

Disagreements about the discount rate to value the impact of the alleged harmful act in periods after the valuation date

As a matter of economics, the appropriate discount rate to convert the economic value of future flows of economic income or loss (e.g., \$100 of business profits in one year) into their economic value as of an earlier date depends on two factors: the time value of money and the risk premium. The time value of money relates to how individuals and companies trade off sure dollars in the future for sure

dollars as of an earlier date, and it is typically measured using a risk-free rate. The risk premium relates to how individuals and companies trade off uncertain dollars for certain dollars. As the name suggests, the risk premium is typically expressed as an increment to the time value of money, such that the discount rate can be calculated as the sum of the time value of money (i.e., the risk-free rate) and the risk premium.

Both factors are important in the determination of discount rates and can have a substantial impact on damages estimates. This section discusses key issues faced by experts when estimating discount rates that adequately measure the time value of money and the risk premium.

Estimating the time value of money

As noted above, the time value of money relates to how individuals and companies trade off dollars in the future for dollars as of an earlier date. The value of money has a temporal dimension for various reasons. First, inflation erodes the value of future money by reducing the basket of goods and services that can be purchased with the same amount of dollars. Second, individuals and companies can reinvest today's dollars in order to obtain a larger number of dollars in the future. Thus, the time value of money on a given date depends on inflation expectations and on available investment opportunities. As those change over time, so does the time value of money. Discount rates used in damages analyses that do not adequately account for the time value of money as of the valuation date may generate unreliable damages estimates.

To estimate the time value of money, experts typically rely on interest rates derived from the prices of safe government debt (in whatever currency damages are being sought and measured). When damages are being measured in U.S. dollars, experts typically use interest rates derived from the prices of U.S. Treasury bills, notes, and bonds of various maturities (ranging from a few days to as many as thirty years). U.S. government debt is typically regarded as risk-free, in the sense that the investor is guaranteed the promised return on their investment if the security is held to its maturity.³⁰

Interest rates derived from longer-term debt tend to differ from interest rates derived from shorter-term debt: Long-term interest rates are typically higher than short-term interest rates in normal economic conditions. Thus, the relevant interest rate to account for the time value of money depends on the horizon of the flows being analyzed. For example, lost profits thirty years into the future should be valued using discount rates that measure the time value of money based on interest rates with a similar horizon.

30. In truth, U.S. government debt is not entirely risk-free, because there is a risk (remote as it may be) that the U.S. government may default.

Estimating the risk premium

When determining the appropriate discount rate to value a flow of economic income or loss, the key source of disagreement among experts tends to be the estimation of the risk premium. Disagreements over this topic also tend to be difficult for the layperson to assess because the academic literature that studies risk premiums is vast and tends to be highly technical. Therefore, particular attention needs to be given to the basic principles surrounding the estimation of risk premiums in order to evaluate experts' positions and their bases.

The risk premium associated with a particular flow of economic income or loss is the compensation investors require for bearing the risks associated with that flow. The risks associated with a particular flow of economic income or loss are typically split into two types: systematic risks and idiosyncratic risks. Systematic risks relate to broader market conditions, that is, risks arising from factors that impact both the flow of economic income or loss at issue and the economy more broadly. Systematic risks affecting a firm include economic downturns, inflation, and other risks that an investor cannot avoid by holding a diversified investment portfolio. In contrast, idiosyncratic risks are risks arising from factors that impact the flow of economic income or loss at issue but do not impact the economy more broadly. Idiosyncratic risks may include whether a venture will succeed, a firm's ability to obtain financing, whether a competitor will develop a similar product, and risks related to the pricing of the product or competitive products, such as the price of inputs.

A basic principle of financial economics is that in order to make risky investments, investors require compensation for risks associated with those investments. Thus, whenever the uncertainty surrounding a flow of economic income or loss is related to overall market conditions, it is necessary to account for that uncertainty in the estimation of the discount rate.

In ideal competitive financial markets, investors demand compensation only for systematic risks, because those are the risks that investors cannot eliminate by holding a well-diversified portfolio of assets. Thus, experts often assume that no risk premium is associated with idiosyncratic risks.

A common approach for determining the risk premium associated with a particular flow of economic income or loss, although not the appropriate one under all circumstances where the calculation of a risk premium is warranted, is part of the Capital Asset Pricing Model (CAPM). The CAPM is a standard model in financial economics to analyze the relation between the risk from investing in a particular asset and the expected rate of return that investors demand for that asset (i.e., its discount rate).³¹ Under the CAPM, the expected return for a given asset

31. See, e.g., Aswath Damodaran, *Damodaran on Valuation: Security Analysis for Investment and Corporate Finance* 31–35 (2d ed. 2006); Richard A. Brealey, Stewart C. Myers & Franklin Allen, *Principles of Corporate Finance* 205–08 (13th ed. 2020).

$(E[R_i])$ is equal to the risk-free rate (R_f) (which compensates investors for the time value of money) plus a risk premium for that specific asset (which compensates investors for taking on the risk from investing in the asset). The risk premium for the asset is equal to the risk premium for the overall market ($E[R_m] - R_f$) multiplied by the asset's *beta*, which is explained below. Mathematically, the CAPM equation for the expected return (or discount rate) on an asset is

$$E[R_i] = R_f + \beta_i \times (E[R_m] - R_f)$$

A key part of applying CAPM to the valuation of a particular asset is to determine the asset's beta (β_i). Beta measures the sensitivity of the asset's returns to changes in returns on the overall market. Assets whose returns are more sensitive to the returns of the overall market tend to decline more when the overall market declines. Because investors require compensation for such declines, assets with higher betas carry higher risk premiums (and, thus, have higher discount rates) than assets with lower betas. To illustrate, for a company whose stock has a beta of 1.5, if the index of value for a representative and broad set of firms³² decreases by 10% over a year, then the value of the company's stock tends to decrease by 15% over that year (1.5 times 10%). Under CAPM, the risk premium for that stock can be calculated as the 1.5 beta multiplied by the historical average risk premium for the overall stock market.³³ The calculation may be presented in the following format:

1. Beta: $\beta_i = 1.5$
2. Market risk premium: $E[R_m] - R_f = 6.0\%$
3. Risk premium (line 1 times line 2): 9.0%

In this example, if the expert estimates the risk-free rate as 2.00%, for example, based on 20-year treasury rates as of the valuation date, then the estimated discount rate will be 11.00% (2.00% plus 9.00%).

Disagreements about the capitalization rate to value the impact of the alleged harmful act in periods before the valuation date

When the direct economic impact of the alleged harmful act is a flow of economic income or loss that occurred (or would have occurred in the but-for world) before the valuation date, it is necessary to determine a capitalization rate that can be used to value the past flow of economic income or loss as of the later valuation date. Such determination often requires careful consideration of the

32. *E.g.*, the S&P 500 index.

33. Experts may disagree on the appropriate methodology to estimate the market risk premium.

specific circumstances faced by the plaintiff that was impacted by the alleged harmful act at the time the direct economic impact occurred.

In some situations, it is natural to see past flows as income that could have been reinvested, and thus, it is necessary to consider the rate of return that investment would have generated. For example, in a matter where a brokerage firm is alleged to have overcharged its client investors for stock transactions over a period of years, investors may claim that the amount of the overcharge in each transaction would have been reinvested into the same stocks and generated an additional return by the valuation date. In other situations, it is natural to see the past flow as a loss that would have to be financed (e.g., from a bank), and it is necessary to consider what rate of return would have been required in order to finance that loss. For example, in a patent infringement case where the patentee's profits were adversely impacted by the infringing company's conduct, the patentee may claim that its lost profits are economically equivalent to an unsecured loan (albeit an involuntary one) to the infringer, and therefore the infringer's borrowing rate should be used to capitalize past lost profits.

The appropriate capitalization rate to value past flows of economic income or loss is not always a matter of expert analysis. Sometimes, it is set by statute or legal precedent (a topic covered in the section titled "Prejudgment Interest" below), and all an expert needs to do, once the direct economic impact of the alleged harmful act in past periods has been estimated, is to capitalize those estimates through the valuation date.

Disagreements about what valuation date to use

An important decision in many analyses of economic damages is the valuation date, that is, the date as of which the value of the direct economic impact of the alleged harmful act over multiple periods will be determined. Sometimes, the appropriate valuation date is set by law. For example, in a breach of contract matter, the proper valuation date is typically the date of the alleged breach. However, when the valuation date is not set by law, expert disputes may arise over the valuation date, given its potential impact on damages.

There are at least two reasons why the valuation date used in a damages analysis can have a material impact on damages. First, the appropriate rates for discounting and capitalization purposes change significantly over time. For example, all else being equal, the appropriate discount rate to value a business's expected future profits will be higher in periods when future inflation is expected to be high. Thus, if there are significant changes in inflation expectations, the damages valuation date can have a significant impact on the damages estimate. Second, the experts' assessment of the direct economic impact of the alleged harmful act across multiple periods depends on the parties' knowledge as of the valuation date, and that knowledge may change over time as the parties'

economic circumstances evolve. For example, in a business valuation case, the value of the business at issue may be very different at a date when a company has no direct competitors versus a date—potentially just a few months later—after competitor firms have entered the market, significantly eroding the profitability of the business at issue.

Other Issues Arising in General in Damages Measurement

Data Validity

Validation of any dataset is critical. The expert needs to have a firm basis for relying on the chosen data. Opposing parties frequently try to impeach damages estimates by challenging the reliability of the data or an expert's validation of the data. This section discusses some key issues that experts consider when evaluating data validity.

Criteria for Determining the Validity of Data

The validity of data is ultimately a matter of judgment. Experts often need to use data that are not mathematically precise because the only relevant available data may be known to contain some errors. Experts generally have an obligation to use data that are as accurate as possible, meaning that the expert has used every practical means to eliminate erroneous information. Experts should also perform cross-checks with other data, to the extent possible, to demonstrate completeness and reliability. Where data are inherently inaccurate because of random influences, validity requires absence of bias or adjustment for bias. Validation of data turns in part on commonsense indicators of accuracy and bias. The following is a list, in rough order of presumptive validity, of data sources often used in damages measurement:

1. Official government publications and databases, such as from the Census Bureau, the Bureau of Labor Statistics, and the Bureau of Economic Analysis
2. Filings with the Securities and Exchange Commission, including a company's audited financial statements
3. A company's accounting and sales records maintained in the normal course of business
4. A company's operating reports prepared for management in the normal course of business
5. Reports issued by securities analysts covering a company

Reference Guide on Estimation of Economic Damages

6. A survey designed by the damages expert with assistance from survey professionals, conforming to established standards of survey design and execution
7. A marketing–research study conforming to established standards for these studies
8. Industry reports and other materials prepared by unaffiliated organizations and consultants
9. Newspaper articles
10. A company’s study of damages from the harmful event, prepared in the normal course of business
11. A company’s study of damages, prepared for litigation

Other factors can alter this presumptive order of validity. When audited financial statements are alleged to be fraudulent, they (or some portions therein) may lose their presumption of validity. The most fully researched articles in the best newspapers have a higher presumption of validity. Some private industry reports are highly reliable. Rules of evidence may also affect when and how these various data sources can be used in expert reports and at trial. However, when an expert’s preferred analysis would rely on data from an opposing party and those data are unavailable, then courts may make allowances for this lack of availability if the expert has made every effort to demonstrate that the data used in the expert’s actual analysis are reasonable.

Quantitative Methods for Validation

One important aspect of validation is to verify that the data are complete. If a separate summary document is available that shows the number of records in the database together with summary statistics such as total amounts paid, completeness is easy to establish. Other methods for establishing completeness include examining serial numbers for records and finding other sources of information about transactions that should be in the data and verifying the presence of all or a sample of them.

Another validation method is to examine the accuracy of a sample of observations in a dataset by comparing the sampled observations to corresponding source documents to ensure the values match. Choosing which specific observations to sample can be done in alternative ways. For example, if the dataset consists of purchasing records, then the expert may examine all of the records for a sample of customers, or the expert may examine all of the records for a sample of transactions of certain types.

Another important aspect of validation is to test the internal integrity of the data. For example, sales proceeds should reflect underlying quantities sold and unit prices. The expert can compare sales proceeds to the product of quantities

and prices. Similarly, an expert may analyze the distribution of values for a particular data field to identify observations with values outside the expected range. For example, a negative value in a field representing a unit price would not make logical sense.

Handling of Missing Data

In dealing with missing data, it is critical to ascertain why the data are missing and to attempt to isolate the extent of the missing data. If only a small fraction of data is missing and the pattern appears to be random, then potentially the issue of the missing data can be disregarded and inferences can be drawn using only the available data.

However, missing data are seldom random. For example, suppose that only 1% of transactions are missing, but the transactions that are missing are large ones accounting for a third of all volume. Validation by summarizing across different characteristics will usually identify missing data that are not random. Such errors might occur if all the missing transactions were submitted in a different format that the program for reading the data does not handle. For example, a manufacturer might record sales to Walmart separately from all other customers.

Identifying and adding the missing data to the database is the best correction for missing data, but this is often not possible. In this case, the expert needs to address the problem in another way. The simplest method is to “gross up” damages to reflect the missing data. For example, if 10% of the transactions are randomly missing, then the expert may correct for the missing transactions by dividing calculated damages by 0.9. This method implicitly assumes that the percentage of missing transactions is known and that the missing and non-missing transactions reflect the same damages on average.

A related approach is to rely on partial, detailed data to measure damages as a fraction of another variable, such as sales. With this approach, other reliable and complete company records such as audited financials may be used to identify the company’s total revenues, and damages are then calculated as the fraction of total sales as calculated above. This approach implicitly assumes that the relationship between the missing variable (i.e., damages) and the related, observable variable (e.g., sales) is stable and that most of the variation in outcomes for the missing variable can be explained by the variation in outcomes for the observable variable. The expert may choose to patch together incomplete data from one source and infer complete, reliable data using other approaches. Such approaches can include using a survey of customers or workers to measure damages per dollar of sales or per dollar of earnings and then applying those ratios to reliable data on total sales or total earnings.

Prejudgment Interest

The law may specify how to calculate interest for losses prior to a verdict on liability, generally termed prejudgment interest.³⁴ The law may exclude prejudgment interest, specify prejudgment interest to be a statutory rate, or exclude compounding.³⁵ Table 1 illustrates these alternatives. With simple uncompounded interest, losses from five years before trial earn five times the specified annual interest rate, so compensation for a \$100 loss from five years ago is \$135 at a 7% annual interest rate. With compound interest, the plaintiff earns interest on past interest. Compensation at 7% interest compounded annually is about \$140 for a loss of \$100 five years before trial. The difference between simple and compound interest becomes much larger if the time from loss to trial is greater or if the interest rate is higher. Because interest is often reinvested to earn further interest, economic analysis generally supports the use of compound interest.

Table 1. Calculations of Prejudgment Interest (in Dollars)

Years of Before Trial	Loss Without Interest	Loss with Compound Interest at 7%	Loss with Simple Uncompounded Interest at 7%
10	\$100	\$197	\$170
9	\$100	\$184	\$163
8	\$100	\$172	\$156
7	\$100	\$161	\$149
6	\$100	\$150	\$142
5	\$100	\$140	\$135
4	\$100	\$131	\$128
3	\$100	\$123	\$121
2	\$100	\$114	\$114
1	\$100	\$107	\$107
0	\$100	\$100	\$100
Total	\$1,100	\$1,578	\$1,485

34. See generally Michael S. Knoll, *A Primer on Prejudgment Interest*, 75 Tex. L. Rev. 293 (1996) (discussing prejudgment interest extensively). See, e.g., *Ford v. Rigidply Rafters, Inc.*, 984 F. Supp. 386, 391–92 (D. Md. 1997) (specifying a method of calculating prejudgment interest in an employment discrimination case to ensure plaintiff is fairly compensated rather than given a windfall); *Arcon/Pac. Ltd. v. Coit*, No. C-81-4264-VRW, 1997 WL 578673, at *1–*2 (N.D. Cal. Sep. 8, 1997) (reviewing supplemental interest calculations and applying California state law to determine the appropriate amount of prejudgment interest to be awarded); *Prestige Cas. Co. v. Michigan Mut. Ins. Co.*, 969 F. Supp. 1029, 1032–34 (E.D. Mich. 1997) (analyzing Michigan state law to determine the appropriate prejudgment interest award).

35. When compounding interest, it is important to use a rate that reflects the periodicity at which interest is compounded (e.g., one should use an annual rate if interest is compounded annually).

Where the law does not prescribe the form of interest for past losses, experts will normally apply a reasonable interest rate to bring those losses forward. The parties may disagree on whether the interest rate should be measured before or after tax. The before-tax interest rate is the normally quoted rate. To calculate the corresponding after-tax rate, one subtracts the amount of income tax the recipient would have to pay on the interest. Thus, the after-tax rate depends on the tax situation of the plaintiff. As an example, if the before-tax interest rate is 9% and the tax rate is 30%, then one would subtract 2.7% (9% times 30%) from the before-tax interest rate of 9% to arrive at an after-tax interest rate of 6.3%.

Even where damages are calculated on a pretax basis, economic considerations suggest that the prejudgment interest rate should be on an after-tax basis: Had a taxpaying plaintiff actually received the lost earnings in the past and invested the earnings at the assumed rate, income tax would have been due on the interest. The plaintiff's accumulated value would be the amount calculated by compounding past losses at the after-tax interest rate.

Where there is economic disparity between the parties, there may be a disagreement about whose interest rate should be used—the borrowing rate of the defendant or the lending rate of the plaintiff or some other rate. There may also be disagreements about adjustment for risk.³⁶

Accounting for Income Taxes

A damages award compensates the plaintiff for lost economic value. In principle, the calculation of compensation should measure the plaintiff's loss after taxes and then calculate the magnitude of the pretax award needed to compensate the plaintiff fully, once taxation of the award is considered. In practice, the tax rates applied to the original loss and to the compensation are frequently the same. When the rates are the same, the two tax adjustments are a wash. In that case, the appropriate pretax compensation is simply the pretax loss, and the damages calculation may be simplified by the omission of tax considerations.

In some damages analyses, explicit consideration of taxes is essential, and disagreements between the parties may arise about these tax issues. If the plaintiff's lost income would have been taxed as a capital gain (at a preferential rate), but the damages award will be taxed as ordinary income, the plaintiff can be expected to include an explicit calculation of the extra compensation needed to

36. See generally James M. Patell, Roman L. Weil & Mark A. Wolfson, *Accumulating Damages in Litigation: The Roles of Uncertainty and Interest Rates*, 11 J. Legal Stud. 341 (1982), <https://doi.org/10.1086/467705> (extensive discussion of interest rates in damages calculations).

make up for the loss of the tax advantage. Sometimes tax considerations are paramount in damages calculations.³⁷

Example: Trustee wrongfully sells Beneficiary's property at full market value. Beneficiary would have owned the property until death and deferred the capital gains tax.

Comment: Damages are the difference between the actual capital gains tax and the present value of the future capital gains tax that would have been paid but for the wrongful sale.

In some cases, the law requires different tax treatment of the claimed loss and compensatory awards. Again, the tax adjustments do not offset each other, and consideration of taxes may be a source of dispute.

Example: Driver injures Victim in a truck accident. A state law provides that awards for personal injury are not taxable, even though the income lost as a result of the injury would have been taxable. Victim calculates damages as lost pretax earnings, but Driver calculates damages as lost earnings after tax.³⁸ Driver argues that the nontaxable award would exceed actual economic loss if it were not adjusted for the taxation of the lost income.

Comment: Under the principle that damages are to restore the plaintiff to the economic equivalent of the plaintiff's position absent the harmful act, it may be recognized that the income to be replaced by the award would have been taxed.³⁹ However, the law in a particular jurisdiction may not allow a jury instruction on the taxability of an award.⁴⁰

37. See generally John H. Derrick, Annotation, *Damages for Breach of Contract as Affected by Income Tax Considerations*, 50 A.L.R. 4th 452 (1986) (discussing a variety of state and federal cases in which courts ruled on the propriety of tax considerations in damage calculations; courts have often been reluctant to award difference in taxes as damages because it is calling for too much speculation).

38. See generally Brian C. Brush & Charles H. Breedon, *A Taxonomy for the Treatment of Taxes in Cases Involving Lost Earnings*, 6 J. Legal Econ. 1 (1996) (discussing four general approaches for treating tax consequences in cases involving lost future earnings or earning capacity based on the economic objective and the tax treatment of the lump sum award).

39. See, e.g., *Myers v. Griffin-Alexander Drilling Co.*, 910 F.2d 1252 (5th Cir. 1990) (holding loss of past earnings between the time of the accident and the trial could not be based on pretax earnings).

40. See generally John E. Theuman, Annotation, *Propriety of Taking Income Tax into Consideration in Fixing Damages in Personal Injury or Death Action*, 16 A.L.R. 4th 589 (1982) (discussing a variety of state and federal cases in which the propriety of jury instructions regarding tax consequences is at issue). See, e.g., *Bussell v. DeWalt Prods. Corp.*, 519 A.2d 1379 (N.J. 1987) (holding that trial court hearing a personal injury case must instruct jury, upon request, that personal injury damages are not subject to state and federal income taxes); *Gorham v. Farmington Motor Inn, Inc.*, 271 A.2d 94 (Conn. 1970) (holding court did not err in refusing to instruct jury that personal injury damages were tax-free).

- Example:* Worker is wrongfully deprived of tax-free fringe benefits by Employer. Under applicable law, the award is taxable. Worker's damages estimate is grossed up by a factor so that the amount of the award, after tax, is sufficient to replace the lost tax-free value.
- Comment:* Again, to achieve the goal of restoring the plaintiff to a position economically equivalent absent the harmful act, an adjustment of this type is appropriate. The adjustment is often called grossing up damages.⁴¹ To accomplish grossing up, divide the lost tax-free value by one minus the tax rate. For example, if the loss is \$100,000 of tax-free income, and the income tax rate is 25%, the award should be \$100,000 divided by 0.75, or \$133,333.

Damages with Multiple Challenged Acts: Disaggregation

Plaintiffs sometimes challenge a number of defendants' acts and offer an estimate of the combined effect of those acts. If the court determines that only some of the challenged acts are illegal, the damages analysis needs to be adjusted to consider only those acts. Ideally, the damages testimony would equip the factfinder to determine damages for any combination of the challenged acts, but that may be tedious. If there are, say, ten challenged acts, it would take more than 1,000 separate studies to determine damages for every possible combination of findings about the unlawfulness of the acts.

There have been several cases where the jury has found partially for the plaintiff, but the jury lacked assistance from the damages experts on how the damages should be calculated for the combination of acts the jury found to be unlawful. Although some of these juries have attempted to resolve these issues, appeals courts have sometimes rejected damages found by juries without supporting expert testimony.⁴²

One solution to this problem is to make the determination of the illegal acts before damages testimony is heard, termed bifurcation of liability and damages.

41. See Cecil D. Quillen, Jr., *Income, Cash, and Lost Profits Damages Awards in Patent Infringement Cases*, 2 Fed. Cir. Bar J. 201, 207 (1992) (discussing the importance of taking tax consequences and cash flows into account when estimating damages).

42. See, e.g., *Litton Sys., Inc. v. Honeywell Inc.*, Case No. CV 90-4823 MRP (Ex), 1996 U.S. Dist. LEXIS 14662, at *13 (C.D. Cal. July 24, 1996) (granting new trial on damages only "[b]ecause there is no rational basis on which the jury could have reduced Litton's 'lump sum' damage estimate to account for Litton's losses attributable to conduct excluded from the jury's consideration"); *Image Tech. Servs., Inc. v. Eastman Kodak Co.*, 125 F.3d 1195, 1224 (9th Cir. 1997), cert. denied, 523 U.S. 1094 (1998) (plaintiffs "must segregate damages attributable to lawful competition from damages attributable to Kodak's monopolizing conduct").

The damages experts can adjust their testimony to consider only the acts found to be illegal.

In some situations, damages are the sum of separate damages for the various illegal acts. For example, there may be one injury in New York and another in Oregon. Then the damages testimony may consider the acts separately, and disaggregation is not challenging.

When the challenged acts have effects that interact, it is not possible to consider damages separately and add up damages for each individual act. This is an area of great confusion. When the harmful acts substitute for each other, the sum of damages attributable to each separately is *less* than their combined effect. As an example, suppose that the defendant has used exclusionary contracts and anticompetitive acquisitions to ruin the plaintiff's business. However, the plaintiff's business could not survive if either the contracts or the acquisitions were found to be legal. Damages for the combination of acts are the value of the business, which would have thrived absent both the contracts and the acquisitions. Now consider damages if only the contracts but not the acquisitions are illegal. In the but-for analysis, the acquisitions are hypothesized to occur because they are not illegal, but not the contracts. But the plaintiff's business cannot function in that but-for situation because the acquisitions alone were sufficient to ruin the business. Hence, damages—the difference in value of the plaintiff's business in the but-for and actual situations—are zero. The same would be true for a separate damages measurement for the acquisitions, with the contracts taken to be legal but not the acquisitions. Thus, the sum of damages for the individual acts is zero, but the damages if both acts are illegal are the value of the business.

When the effects of the challenged conduct are complementary, the sum of damages for each type of conduct by itself will be *more* than damages for all types of conduct together. For example, suppose a party claims that a contract is exclusionary based on the combined effect of the contract's duration and its liquidated damages clause that includes an improper penalty provision. The actual amount of the penalty would cause little exclusion if the duration were brief, but substantial exclusion if the duration were long. Similarly, the actual duration of the contract would cause little exclusion if the penalty were small but substantial exclusion if the penalty were large. A damages analysis for the penalty provision in isolation compares but-for—without the penalty provision but with long duration—to actual, where both the penalty provision and long duration are in effect, and finds that damages are large. Similarly, a damages estimate for the duration in isolation gives large damages. The sum of the two estimates is nearly double the damages from the combined use of both provisions.

A request that the damages expert disaggregate damages among each of the individual challenged acts is therefore complex and will not necessarily lead to individual estimates that will add up to the estimate based on all of the challenged acts. In principle, a separate damages analysis—with its own carefully

specified but-for scenario and analysis—needs to be done for every possible combination of illegal acts.⁴³

Example: Hospital challenges Glove Maker for illegally obtaining market power through the use of long-term contracts and the use of a discount program that gives discounts to consortiums of hospitals if they purchase exclusively from Glove Maker. The jury finds that Glove Maker has attempted to monopolize the market with its discount programs but that the long-term contracts were legal because of efficiencies. Hospital states that its damages are the same as in the case in which both acts were unlawful because either act was sufficient to achieve the observed level of market power. Glove Maker argues that damages are zero because the lawful long-term contracts would have been enough to allow it to dominate the market.

Comment: The appropriate damages analysis is based on a careful new comparison of the market with and without the discount program. The but-for analysis should include the presence of the long-term contracts because they were found to be legal.

Disputes About Whether the Plaintiff Is Entitled to All the Damages

When the plaintiff is in some sense a conduit to other parties, the defendant may argue that the plaintiff is entitled to only those damages that it would have retained in the but-for scenario. In the following example, a regulated utility is arguably a conduit to the ratepayers:

Example: Generator Maker overcharges Utility. Generator Maker argues that the overcharge would have been part of Utility's rate base, and so Utility's regulator had set higher prices because of the overcharge. Utility, therefore, did not lose anything from the overcharge. Instead, the ratepayers paid the overcharge. Utility argues that it stands in for all the ratepayers and that the damages

43. Order Denying Motion for Class Certification, *Reitman v. Champion Petfoods United States, Inc.*, No. 18-cv-1736-DOC-JPRx, 2019 U.S. Dist. LEXIS 221941 (C.D. Cal. Oct. 30, 2019). In denying class certification the court found that “the [Plaintiffs’ damages] model does not test whether the corrective statements have a different impact on consumers in particular combinations. Therefore, if any theory of liability is disposed of, the survey does not provide a reliable way to determine the difference in value based on the remaining theories of liability because the expert did not test the impact of particular combinations of statements. For these reasons, Plaintiffs have not ‘show[n] that such damages can be determined without excessive difficulty and attributed to their theory of liability.’” *Id.* at *32 (citing *JustFilm, Inc. v. Buono*, 847 F.3d 1108, 1121 (9th Cir. 2017)).

award will accrue to the ratepayers by the same principle—the regulator will set lower rates because the award will count as revenue for rate-making purposes.

Comment: In addition to the legal issue of whether Utility does stand in for ratepayers, there are two factual issues: Was the overcharge actually passed on to ratepayers? Will the award be passed on to ratepayers?

Similar issues can arise in the context of employment law.

Example: Plaintiff Sales Representative sues for wrongful denial of a commission. Sales Representative has subcontracted with another individual to do the actual selling and pays a portion of any commission to that individual as compensation. The subcontractor is not a party to the suit. Defendant Manufacturer argues that damages should be Sales Representative's lost profit measured as the commission less costs, including the payout to the subcontractor. Sales Representative argues that they are entitled to the entire commission.

Comment: Given that the subcontractor is not a plaintiff, and Sales Representative avoided the subcontractor's commission, the literal application of standard principles for damages measurement would appear to call for the lost-profit measure. The subcontractor, however, may be able to claim their share of the damages award. In that case, damages would equal the entire lost commission, so that, after paying off the subcontractor, Sales Representative receives exactly what they would have received absent the breach.

Limitations on Damages

Uncertainty and Speculation in Damages Estimates

Generally, a plaintiff may not recover damages beyond an amount proven with reasonable certainty.⁴⁴ This rule permits damages estimates that are not mathematically certain but excludes those that are speculative.⁴⁵ Failure to prove damages to a reasonable certainty is a common defense.⁴⁶

44. See, e.g., Restatement (Second) of Contracts § 352 (1981) ("Damages are not recoverable for loss beyond an amount that the evidence permits to be established with reasonable certainty.").

45. Restatement (Second) of Contracts § 352 cmt. a states, in pertinent part: "Damages need not be calculable with mathematical accuracy and are often at best approximate."

46. See, e.g., *Cole v. Homier Distrib. Co.*, 599 F.3d 856, 866 (8th Cir. 2010) (expert testimony on lost profits excluded under *Daubert* standard because it "failed to rise above the level of speculation."). See also *Webb v. Braswell*, 930 So. 2d 387 (Miss. 2006). In *Webb*, the plaintiff's expert

However, defining what constitutes reasonable certainty or speculation in a damages analysis may be difficult. The exclusion of damages on grounds of excessive uncertainty regarding the amount of damages may result in an award of zero damages—even when it is likely that the plaintiff suffered significant damages, though the actual amount is quite uncertain.

Traditionally, damages are calculated without reference to uncertainty about outcomes in the but-for scenario. Outcomes are taken as actually occurring if they are the expected outcome and as not occurring if they are not expected to occur. This approach may overcompensate some plaintiffs and undercompensate others. For example, suppose that a drug company was deprived of the opportunity to bring to market a drug that had a 90% chance of receiving Food and Drug Administration (FDA) approval, at a profit of \$2 billion, and a 10% chance of not receiving FDA approval, with losses of \$1 billion. The court may treat 90% as near enough to certainty and ignore the 10% risk of failure and award damages of \$2 billion.

By contrast, economists quantify losses from uncertain outcomes in terms of expected values, where the value in each outcome is weighted by its probability. Under that approach, economic losses in the above example would be calculated as the foregone \$2 billion potential profit assuming FDA approval times 90% plus the \$1 billion potential loss times 10%, or $(0.9) \times (\$2 \text{ billion}) + (0.1) \times (-\$1 \text{ billion}) = \$1.7 \text{ billion}$. Thus, the plaintiff would be overcompensated by \$300 million under the traditional approach that ignored the small probability of failure.

Now suppose the drug has only a 40% chance of FDA approval with the same economic payoffs. Although the court may decide to award the plaintiff no damages on grounds of uncertainty and speculation, the economic loss would be calculated as $(0.4) \times (\$2 \text{ billion}) + (0.6) \times (-\$1 \text{ billion}) = \$200 \text{ million}$. This issue can arise in any situation involving businesses or products whose future profitability is highly uncertain due to an unproven track record of profitability.

Damages That Are Too Remote

A second limitation on damages is that a plaintiff may not recover damages that are too remote. In tort cases, this restriction is expressed in terms of proximate cause,⁴⁷ which often is equivalent to reasonable foreseeability. In contract cases,

sought to testify as to future damages resulting from unplanted crops, without establishing that the crops would have been profitable. The court excluded the testimony based on Mississippi's adoption of the *Daubert* standard, stating that “damages for breach of contract must be proven with reasonable certainty and not based merely on speculation and conjecture.” *Id.* at 397.

47. See William L. Prosser, *Palsgraf Revisited*, 52 Mich. L. Rev. 1 (1953); Osborne M. Reynolds, Jr., *Limits on Negligence Liability: Palsgraf at 50*, 32 Okla. L. Rev. 63 (1979).

the limitation is similarly embodied in the idea of reasonable foreseeability—a party may not recover damages that were not reasonably foreseeable by the parties at the time of the agreement.⁴⁸ Generally, a party is liable for what are known as direct or general damages—those damages that arise naturally from the breach itself. A defendant may also be liable for consequential or special damages—damages apart from those arising naturally from the breach—if such damages were reasonably foreseeable at the time of the agreement.⁴⁹

Mitigation of Losses

A third limitation on damages is that a party may not recover for losses it could have avoided; this limitation is often expressed by stating that the injured party has a duty to mitigate, or lessen, its damages. The economic justification for the mitigation rule is that the injured party should not cause economic waste by needlessly increasing its losses.⁵⁰

For example, in a dispute about mitigation, a defendant may propose that any damages estimate should include an offset for earnings the plaintiff should have achieved, under proper mitigation, rather than actual earnings. As another example, the defendant might presume that the plaintiff could mitigate by locating another source of supply in the event of a breach of a supply agreement. Damages are limited to the difference between the contract price and the current market price in that situation.

For personal injuries, the issue of mitigation often arises because the defendant believes that the plaintiff's failure to work after the injury is a withdrawal from the labor force or retirement rather than the result of the injury.⁵¹ For commercial torts, mitigation issues can be more subtle. Where the plaintiff believes that the harmful act destroyed a company, the defendant may argue that the company could have been put back together and earned profit, possibly in a different line of business.⁵² The defendant will then treat the hypothetical profits as an offset to damages.⁵³

Alternatively, where the plaintiff continues to operate the business after the harmful act and includes subsequent losses in damages, the defendant may argue that the proper mitigation was to shut down after the harmful act.⁵⁴

48. See E. Allan Farnsworth, *Legal Remedies for Breach of Contract*, 70 Colum. L. Rev. 1145, 1199–1210.

49. *Id.*

50. See *id.* at 1183–84.

51. See William T. Paulk, *Mitigation Through Employment in Personal Injury Cases: The Application of the "Reasonable" Standard and the Wealth Effects of Remedies*, 58 Ala. L. Rev. 647–64 (2007).

52. See *Seahorse Marine Supplies Inc. v. Puerto Rico Sun Oil Co.*, 295 F.3d 68, 84–85 (1st Cir. 2002).

53. *Id.* at 84.

54. *Id.* at 85. See also *In re First New England Dental Ctrs., Inc.*, 291 B.R. 229, 240 (D. Mass. 2003).

Example: Franchisee Soil Tester starts up a business based on Franchiser's proprietary technology, which Franchiser represents as meeting government standards. During the startup phase, Franchiser notifies Soil Tester that the technology has failed. Soil Tester continues to develop the business but sues Franchiser for profits it would have made from successful technology. Franchiser calculates much lower damages on the theory that Soil Tester should have mitigated by terminating the startup.

Disagreements about mitigation may be hidden within the frameworks of the plaintiff's and the defendant's damages studies.

Example: Defendant Board Maker has breached an agreement to supply circuit boards. Plaintiff Computer Maker's damages study is based on the loss of profits on the computers to be made from the circuit boards. Board Maker's damages study is based on the difference between the contract price for the boards and the market price at the time of the breach.

Comment: There is an implicit disagreement about Computer Maker's duty to mitigate by locating alternative sources for the boards not supplied by the defendant. The Uniform Commercial Code spells out the principles for resolving these legal issues under the contracts it governs.⁵⁵

Damages That May Exceed the Cost of Avoidance

An important consideration in capping damages may be the costs of steps that the plaintiff could have taken that would have eliminated damages. This argument is closely related to mitigation but has an important difference: The defendant may argue that the plaintiff's failure to undertake a costly step that would have avoided losses was reasonable, but that the failure to take that step shows that the plaintiff knew that damages were much smaller than its later damages claim.

55. See, e.g., *Aircraft Guar. Corp. v. Strato-Lift, Inc.*, 991 F. Supp. 735, 738–39 (E.D. Pa. 1998) (Mem.) (finding that according to the Uniform Commercial Code, plaintiff-buyer had a duty to mitigate if the duty was reasonable in light of all the facts and circumstances, but that failure to mitigate does not preclude recovery); *S.J. Groves & Sons Co. v. Warner Co.*, 576 F.2d 524 (3d Cir. 1978) (holding that the duty to mitigate is a tool to lessen plaintiff's recovery and is a question of fact); *Thomas Creek Lumber & Log Co. v. United States*, 36 Fed. Cl. 220 (1996) (finding that under federal common law, the U.S. government had a duty to mitigate in breach-of-contract cases).

Example: Insurance Company suffered a business interruption because a fire made its offices unusable for a period of time. Insurance Company's damages claim for \$10 million includes not only the lost business until the offices were usable but also damages for permanent loss of business from customers who found other sources during the period the offices were unusable. Defendant argues that the plaintiff's failure to relocate to temporary quarters shows that their losses were less than the \$350,000 cost of that relocation.

Comment: Defendant's argument has the unstated premise that Insurance Company could have carried on its business and avoided any of its later losses by relocating. Insurance Company will likely argue that a decision not to relocate was commercially appropriate because relocation would not have avoided much of the lost business.

The Effect of a Liquidated Damages Clause

In addition to legally imposed limitations on damages, the parties themselves may have agreed to impose limits on damages, should a dispute arise. Such clauses are common in many types of agreements. Once litigation has begun, the parties may dispute whether these provisions are legally enforceable. The law may limit enforcement of liquidated damages provisions to those that bear a reasonable relation to the actual damages. In particular, the defendant may attack the amount of liquidated damages as an unenforceable penalty. The parties may disagree on whether the harmful event falls within the class intended by the contract provision.

Changes in economic conditions may be an important source of disagreement about the reasonableness of a liquidated damages provision. One party may seek to overturn a liquidated damages provision on the grounds that new conditions make it unreasonable.

Example: Scrap Iron Supplier breaches supply agreement and pays only the specified liquidated damages. Buyer seeks to set aside the liquidated damages provision because the price of scrap iron has risen, and the liquidated damages are a small fraction of actual damages under the expectation principle.

Comment: There may be conflict between the date for judging the reasonableness of a liquidated damages provision and the date for measuring expectation damages, as in this example. Generally, the date for evaluating the reasonableness of liquidated damages is the date the contract is made. In contrast, the date for measuring

expectation damages is the date of the breach. The conflict may be resolved by the substantive law of the jurisdiction.

Damages in Class Actions

Class Certification

The U.S. Supreme Court's decision in *Comcast Corp. v. Behrend* states that “rigorous analysis” is required to establish if the model proposed by plaintiffs in a class action is capable of measuring damages on a class-wide basis.⁵⁶ Consistent with this requirement, courts commonly analyze damages models as part of their decision whether to certify a class.⁵⁷

A key question often analyzed by courts is whether the proposed method of measuring class-wide damages isolates the impact of the theory of harm alleged. In *Comcast*, the Supreme Court found that the lower court erred in certifying a class of cable television subscribers because the plaintiffs had failed to show that the proposed damages model adequately translated the legal theory of the harmful event into an analysis of the economic impact of that event. Specifically, the plaintiffs proposed a damages model that estimated the combined impact of four proposed theories of antitrust impact but failed to estimate damages specific to each theory. Because the lower court eventually dismissed all but one of the plaintiffs' theories of harm and the plaintiffs' expert acknowledged that his model did not isolate damages resulting from any one of the plaintiff's theories of harm, the Supreme Court found that the proposed damages model was inadequate in that it was unable to isolate the harm resulting from the single surviving theory.⁵⁸

Since *Comcast*, many courts have emphasized the importance of having a damages model that ties directly to the theory of harm. For example, in *Reitman v. Champion*, a matter involving alleged misrepresentations on the packaging of dog food, the court denied class certification because the plaintiffs' damages

56. 569 U.S. 27, 35 (2013).

57. See, e.g., Memorandum of Law & Order, *Johnson v. Evangelical Lutheran Church in Am.*, No. 11–00023 (MJD/LIB), 2013 U.S. Dist. LEXIS 41944 (D. Minn. Mar. 26, 2013) (court evaluated if Board of Pensions of the Evangelical Lutheran Church in America breached its fiduciary duties by reducing annuity payments and rejected class certification on the basis that the board actions helped rather than harmed the majority of putative class members); Memorandum Opinion and Order, *Indiana Pub. Requirement Sys. v. AAC Holdings, Inc.*, No. 3:19-cv-00407, 2023 U.S. Dist. LEXIS 31022 (M.D. Tenn. Feb. 24, 2023) (The court evaluated whether plaintiff had put forth a reliable method to estimate damages consistent with the alleged misrepresentations. The court rejected plaintiff's damages model with respect to plaintiff's materialization-of-the-risk theory because it did not factor in the risk on which the theory was based. The court accepted plaintiff's damages model otherwise.).

58. *Comcast*, 569 U.S. at 38.

methodology tested the effects of statements created by plaintiffs to correct the alleged misrepresentations, and not the effects of the alleged misrepresentations on consumers.⁵⁹ Similarly, in *Freitas v. Cricket Wireless*, a case involving the sale of 4G-capable phones in markets where the defendant did not actually provide 4G coverage, the court denied class certification because plaintiffs simply measured the price difference between various 3G and 4G phones, but failed to account for other differences in the phones, such as the brand and features of the phones.⁶⁰

Another question that courts commonly analyze in connection with the requirements for class certification is whether the damages model proposed by the plaintiffs results in the need for individualized inquiry.⁶¹ Courts have usually accepted damages models that offer a common or formulaic method to estimate damages even if those models result in individualized damages calculations.⁶² However, class certification may be rejected in cases where computing individual damages is deemed overly complex or fact-specific.⁶³

A third common question is whether the proposed damages methodology is sufficiently developed for the court to assess if the model is capable of measuring class-wide damages. Courts have rejected models that fail to explain how important variables will be accounted for,⁶⁴ that do not plausibly explain how the consequences of the allegations will be isolated,⁶⁵ or that are “vague, indefinite, and unspecific.”⁶⁶ At the same time, some courts have accepted damages models that have not been implemented but merely proposed at an adequate level of detail.⁶⁷

59. The court also indicated that “if any theory of liability is disposed of, [the plaintiffs’ damages model] does not provide a reliable way to determine the difference in value based on the remaining theories of liability because the expert did not test the impact of particular combinations of statements.” See Order Denying Motion for Class Certification, *Reitman v. Champion Petfoods USA, Inc.*, No. 18-cv-1736-DOC-JPRx, 2019 U.S. Dist. LEXIS 221941 (C.D. Cal. Oct. 30, 2019).

60. *Freitas v. Cricket Wireless, LLC*, No. C 19-07270 WHA, 2022 WL 3018061, at *6 (N.D. Cal. July 29, 2022).

61. See, e.g., *Ludlow v. BP, P.L.C.*, 800 F.3d 674 (5th Cir. 2015).

62. See, e.g., *In re Whirlpool Corp.*, 722 F.3d 838 (6th Cir. 2013); *Leyva v. Medline Indus., Inc.*, 716 F.3d 510 (9th Cir. 2013).

63. In *Behrend v. Comcast Corp.*, the court reiterated that class-certification damages should not require “labyrinthine individual calculations.” 655 F.3d 182, 206 (3d Cir. 2011). See also *Brown v. Electrolux Home Prods., Inc.*, 817 F.3d 1225 (11th Cir. 2016) (stating “individual damages defeat predominance if computing them ‘will be so complex, fact-specific, and difficult that the burden on the court system would be simply intolerable.’”) (citation omitted).

64. *Bruton v. Gerber Prods. Co.*, No. 12-cv-02412-LHK, 2018 U.S. Dist. LEXIS 30814, at *34 (N.D. Cal. Feb. 13, 2018).

65. *Ang v. Bimbo Bakeries USA, Inc.*, No. 13-cv-01196-HSG, 2018 U.S. Dist. LEXIS 149395, at *2 (N.D. Cal. Aug. 31, 2018).

66. See also *Ohio Pub. Emps. Ret. Sys. v. Fed. Home Loan Mortg. Corp.*, No. 4:08CV0160, 2018 WL 3861840 (N.D. Ohio Aug. 14, 2018).

67. See, e.g., *In re Scotts EZ Seed Litig.*, 304 F.R.D. 397 (S.D.N.Y. 2015). See also *Goldemberg v. Johnson & Johnson Consumer Cos.*, 317 F.R.D. 374 (S.D.N.Y. 2016).

A court operating under the rule that class certification requires a fully developed damages quantification may need to grant discovery prior to class certification to support the class's damages analysis and the defendant's opposition.

Finally, as in other types of litigation, courts have evaluated damages models proposed in class actions with respect to their relevance and reliability. Many courts have rejected damages models proposed in class actions because they fail to meet the standard of relevance and reliability set by *Daubert*.⁶⁸

Class-Wide Damages

In many class actions, the damages expert measures and testifies to class-wide damages using methods discussed elsewhere in this reference guide. At a high level, the damages method offered by the expert may take any of a number of forms. Among other possibilities, it may involve a uniform damages amount per class member, or per instance of harm, multiplied by the total number of class members or instances of harm; it may involve a uniform damages amount per time period (e.g., workweek) multiplied by the number of at-issue time periods; it may involve a uniform damages percentage (e.g., an overpayment percentage) that is applied to total sales; or it may involve the application of a common method (e.g., a formula) to calculate differing damages amounts for different class members based on various inputs.

In class actions involving securities, damages experts are usually required to estimate per-share damages. Class-wide damages are typically calculated in a subsequent claims process where individual shareholders submit claims, along with evidence of their trades in the securities at issue.⁶⁹

Finally, courts do not typically require class-wide damages to be measured with certainty. As the Supreme Court stated, "Calculations need not be exact."⁷⁰ While mere speculation or guess is not sufficient, courts have held that it is enough if class-wide damages are demonstrated "as a matter of just and reasonable inference."⁷¹

68. See, e.g., *IBEW Local 90 Pension Fund v. Deutsche Bank AG*, No. 11 Civ. 4209 KBF, 2013 WL 5815472 (S.D.N.Y. Oct. 29, 2013); *In re Blood Reagents Antitrust Litig.*, 783 F.3d 183 (3d Cir. 2015). Note, however, that there is disagreement across district courts about whether the *Daubert* standard should be applied to evidence presented at the class-certification stage. See, e.g., *Cox v. Zurn Pex, Plumbing Prods. Liab. Litig.*, 644 F.3d 604 (8th Cir. 2011); *Sali v. Corona Reg'l Med. Ctr.*, 909 F.3d 996, 1005–06 (9th Cir. 2018).

69. See, e.g., *Order Regarding Plaintiffs' Motions for Prejudgment Interest, Approval of Notice of Verdict and Claims Administration Procedure, and Unsealing Documents* (Dkt. Nos. 745, 748, 752), *Hsingching Hsu v. Puma Biotechnology, Inc.*, No. SACV 15–00865 AG (SHKx), 2019 U.S. Dist. LEXIS 154278 (C.D. Cal. Sep. 9, 2019).

70. *Comcast Corp. v. Behrend*, 569 U.S. 27, 35 (2013).

71. See, e.g., *Behrend v. Comcast Corp.*, 655 F.3d 182, 203–04 (3d Cir. 2011) (quoting *Story Parchment Co. v. Paterson Parchment Paper Co.*, 282 U.S. 555, 563 (1931)). See also *In re MyFord Touch Consumer Litig.*, 291 F. Supp. 3d 936 (N.D. Cal. 2018).

Damages of Individual Class Members

Damages experts sometimes have a role in the process of disbursing funds from a verdict or settlement to individual class members. For example, an expert can develop software that measures individual damages based on evidence supplied by individuals through a claims-processing facility. In *Millsap v. McDonnell Douglas*, more than 1,000 class members—persons affected by the defendant’s challenged layoffs—completed sworn questionnaires describing their post-layoff experiences.⁷² McDonnell Douglas supplied additional information from its employee records. The class’s damages expert used standard methods for valuing claims for lost earnings to calculate estimates for each class member. Because the settlement reflected a compromise among the parties on a number of disputes about the law and about the facts underlying the layoffs, the total cash from the settlement was less than the sum of these estimates, and class members received a fraction of the amount indicated by the damages model.

Damages as a Basis to Evaluate the Fairness of a Proposed Settlement

Class-wide damages measures have a key role in resolving class-action cases, because courts refer to them in determining the fairness of proposed settlements. The court’s careful review of the benefits proposed for the class is essential because the interests of the class’s counsel may not be aligned with those of the class members with respect to settlement.

Example: Lender required excessive escrow deposits for property taxes from a class of mortgage borrowers, although the excess was repaid at the end of each year. Under the terms of the proposed settlement negotiated between Lender’s lawyers and those for the class, the excess is to be refunded to the class members immediately, with 30% of that amount paid to class counsel as fees.

Comment: This settlement is unreasonable and leaves the class worse off than they were under the excessive escrows. The loss to the class from placing funds in an escrow is the foregone interest on the amount in the escrow, which would likely be no more than 10% of the excess amount of the escrow. By granting 30% of the refund as fees to class counsel, the class members are at least 20% worse off than they would be if the excess were repaid with a delay.

72. No. 94-cv-633-H(M), 2003 WL 21277124 (N.D. Okla. May 28, 2003).

Damages Issues in Selected Types of Cases

Estimation of Economic Damages in Consumer Product Liability and Misrepresentation Cases

Consider consumer class actions in which a product is alleged to have a particular defect that the plaintiff claims should have been disclosed to consumers, or in which a product is alleged to have been misrepresented to consumers before they purchased it. In these types of cases, a common damages claim is that consumers would not have bought the product, or would have paid less for it, had they known about the alleged defect or had the product not been misrepresented.⁷³ The following sections discuss issues that often appear in the calculation of damages associated with these claims.

Defining the Plaintiff's Economic Position in the Actual and But-For Worlds

As explained earlier in this reference guide, the role of the damages expert is to translate the plaintiff's theory of harm into an objective model for damages. In the case of overpayment damages in a product liability or misrepresentation case, the plaintiff's actual economic position is usually defined as the price paid for the allegedly defective product at a then-current market price without the benefit of appropriate disclosures from the manufacturer.⁷⁴ Plaintiffs commonly offer invoices, sales receipts, or other retail sales data as evidence of purchase and of the prices paid. Disputes may arise if the plaintiff cannot offer direct evidence of the quantity of products purchased or the price paid for them (for example, if the manufacturer has data only on sales to distributors), or if the price paid was different across buyers due to individual discounts or negotiations.

The plaintiff's position in the but-for world is generally more complicated to define. In cases where specific statements made by the product manufacturer allegedly misrepresent the product (for example, by including incorrect information on

73. See, e.g., *In re MyFord Touch Consumer Litig.*, No. 13-cv-3072-EMC (N.D. Cal. Oct. 13, 2015); First Amended Class Action Complaint for Damages, *Zakaria v. Gerber Prods. Co.*, No. 2:15-cv-0200-JAK (Ex), 2015 WL 1085581 (C.D. Cal. Feb. 27, 2015). Other common damages claims are that the manufacturer profited unfairly from the sale of the defective product and that the plaintiffs suffered economic losses in their attempt to repair the defect. Issues arising from these claims can be analyzed using the frameworks provided in the preceding sections.

74. See, e.g., *Hendricks v. Starkist Co.*, No. 13-cv-00729-HSG, 2016 U.S. Dist. LEXIS 134872 (N.D. Cal. Sep. 29, 2016); *Elkies v. Johnson & Johnson Servs.*, No. 2:17-cv-07320-GW-JEMx, 2020 U.S. Dist. LEXIS 256341 (C.D. Cal. June 22, 2020).

the product label), a standard approach is to assume that in the but-for world such misrepresentations would have not occurred (for example, assuming that the label would not include the allegedly misleading information).⁷⁵ Cases in which the plaintiff alleges that the manufacturer omitted relevant information that would have been material for consumers can be more challenging for the damages expert because the characteristics of the disclosure (e.g., wording, size) in the hypothetical but-for world can have significant implications for how consumers react to it and the market price they pay.⁷⁶ Moreover, if the alleged defect only manifests in some products with a given probability,⁷⁷ the specific way in which information about that probability is conveyed to consumers can also have implications for how they react to it and the market price they pay.⁷⁸

Determining the But-For Price

In some cases, the market price of the product, adjusted to reflect the effect of the allegations, may be an appropriate estimate of the but-for price. For example, if the manufacturer of the product later added a warning to the product that corrected the alleged omissions in a way that is consistent with plaintiffs' allegations, the market price of the product in the period after the warning was added may be an appropriate measure of the price of the product in the but-for world. In this case, damages experts must account for other factors beyond the addition of the warning that may have affected the price of the product over the same period in order to isolate the effect of the alleged misconduct. For example, changes in supply or demand conditions after a warning label was added to a product may lead to changes in the price of the product that are unrelated to the alleged misconduct.

Damages experts may also rely on data from ostensibly similar products or assets to account for the effect of confounding factors. Empirical methods such as regression analysis can use information from similar products to account for confounding factors such as market-wide trends, geographic variation, and

75. See, e.g., *Jones v. ConAgra Foods Inc.*, No. C 12–01633 CRB, 2014 U.S. Dist. LEXIS 81292 (N.D. Cal. June 13, 2014); *Larsen v. Vizio, Inc.*, No. 8:14-cv-01865-CJC-JCG (C.D. Cal. May 5, 2015).

76. For example, a warning label with negative language about the product could lead to a very different outcome compared to neutral language. See, e.g., Christopher M. Heaps & Tracy B. Henley, *Language Matters: Wording Considerations in Hazard Perception and Warning Comprehension*, 133 J. Psych. 341 (1999), <https://doi.org/10.1080/00223989909599747>.

77. See, e.g., *In re Takata Airbag Prod. Liab. Litig.*, 1:14-cv-24009-FAM (1787) (S.D. Fla. Apr. 24, 2021).

78. Daniel Kahneman & Amos Tversky, *Prospect Theory: An Analysis of Decision Under Risk*, 47 *Econometrica* 263 (1979), <https://doi.org/10.2307/1914185>.

variation in product features, among others.⁷⁹ For these types of analyses, experts need to demonstrate that the benchmarks used are sufficiently similar to the product at issue and that the model adequately captures the most important features of a product.⁸⁰ Experts also need to show that the prices of benchmark products are affected by similar factors as the price of the product at issue, other than the alleged misconduct (i.e., the prices have “parallel trends”).⁸¹

Some damages experts in product liability and misrepresentation class actions have not relied on observed market prices but instead have relied on stated preference models, which use surveys to try to assess the impact of the alleged omissions or misrepresentations on consumer demand.⁸² Some plaintiff experts have relied directly on the results from stated preference models to estimate per-unit overpayment damages, while other experts have combined estimates of impact on consumer demand with models or assumptions about the supply side of the market to generate estimates of but-for prices.

Conjoint Analysis and Supply-Side Considerations

A stated preference model commonly used by experts for plaintiffs is based on a methodology called conjoint analysis.⁸³ Conjoint analysis is a survey-based methodology used to analyze consumer preferences for products and product features.⁸⁴ Conjoint analysis assumes that the subjective value consumers perceive from a product is the sum of the benefits they derive from individual product features or attributes.⁸⁵ In a “choice-based” conjoint analysis (the type most often

79. See, e.g., Mark Campbell et al., A.B.A., *Damages in Consumer Class Actions*, in A Practitioner’s Guide to Class Actions (Marcy Hogan Greer & Amir Nassihi eds., 3d ed. 2021).

80. See, e.g., Kelly C. Bishop et al., *Best Practices for Using Hedonic Property Value Models to Measure Willingness to Pay for Environmental Quality*, 14 Rev. Env’t Econ. & Pol’y 260 (2020), <https://doi.org/10.1093/reep/reaa001>.

81. Joshua D. Angrist & Jörn-Steffen Pischke, *Mostly Harmless Econometrics: An Empiricist’s Companion* (2008).

82. See the section titled “Realistic Modeling of Preferences, Incentives, and Constraints of Buyers and Sellers” above.

83. For applications in the automotive industry, see, e.g., *In re Gen. Motors LLC Ignition Switch Litig.*, No.1:14-MD-02543-JMF (S.D.N.Y.); *In re Chrysler-Dodge-Jeep Ecodiesel Mktg., Sales Pracs. & Prods. Liab. Litig.*, No. 3:17-md-02777- EMC (N.D. Cal.). For applications relating to food products, see, e.g., *Zakaria v. Gerber Prods. Co.*, No. 2:15-cv-0200-JAK, 2015 WL 1085581 (C.D. Cal.); *Colangelo v. Champion Petfoods USA, Inc.*, No. 6:18-cv-01228-LEK-DEP, 2020 WL 777462 (N.D.N.Y.). For an application in the consumer electronics industry, see *Davidson v. Apple, Inc.*, No. 5:16-cv-4942-LHK, 2018 WL 10716622 (N.D. Cal.).

84. Vithala R. Rao, *Applied Conjoint Analysis* (2014).

85. For example, an expert could assume that the subjective value that a consumer obtains from buying a TV is equal to the value they derive from the TV having a 55-inch screen, plus the value of having an LED screen, plus the value of having smart features, minus the value sacrificed by paying the TV’s price.

used in litigation), the researcher uses a survey to ask consumers to choose which product they prefer among hypothetical products that vary across product attributes. The researcher then analyzes respondents' choices using statistical methods to generate an estimate of the monetary value that a consumer places on different attributes (also known as willingness to pay).⁸⁶ A conjoint analysis used in a product liability case will include one or more attributes that relate to the plaintiffs' allegations. For example, if the plaintiffs claim that the frame-refresh rate of a TV was overstated, the conjoint survey could include some hypothetical product options with the frame-refresh rate advertised to consumers, and other options with the (allegedly lower) correct frame-refresh rate. Based on respondents' answers (and the assumptions underlying the conjoint analysis methodology), an expert may be able to estimate the price that a consumer would be willing to pay for a TV with the advertised frame-refresh rate over a TV with the lower refresh rate.

An expert relying on a conjoint analysis should demonstrate that it provides a sufficiently realistic representation of the preferences of the buyers of the product. This may involve demonstrating that the model includes an appropriate set of attributes that define the demand for the product, that respondents are able to choose not to buy any of the options shown to them (known as the no-purchase option), that the predictions generated by the model are reliable, and that the calculation of respondents' willingness to pay is consistent with the assumptions underlying the conjoint methodology.⁸⁷ When implemented correctly and using a representative sample of respondents, an expert could, for example, rely on conjoint analysis to estimate the demand for TVs with and without the misrepresentation at issue.⁸⁸

As explained in the section titled "Issues Arising in Connection with the Use of Models to Simulate the But-For Price" above, because market prices are determined by the interaction of demand and supply, consumer-survey-based methods like conjoint analysis by themselves are insufficient to estimate the market price of the product at issue in the but-for world. Experts using conjoint analysis have relied on at least two different assumptions to attempt to account for supply-side factors. Some experts have assumed that the quantity supplied in the but-for world is the same as the quantity supplied in the actual world.⁸⁹ The but-for price obtained with this method will be the price necessary to sell the

86. See, e.g., Moshe Ben-Akiva, Daniel McFadden & Kenneth Train, *Foundations of Stated Preference Elicitation: Consumer Behavior and Choice-Based Conjoint Analysis*, 10 *Found. & Trends Econometrics* 1 (2019), <http://dx.doi.org/10.1561/08000000036>.

87. John R. Hauser & Vithala R. Rao, *Conjoint Analysis, Related Modeling, and Applications*, in *Marketing Research and Modeling: Progress and Prospects* 141–68 (Yoram (Jerry) Wind & Paul E. Green eds., 2004). See also Rao, *supra* note 84.

88. See Ben-Akiva, *supra* note 86.

89. See, e.g., *In re Dial Complete Mktg. & Sales Pracs. Litig.*, 320 F.R.D. 326 (D.N.H. 2017); *In re NJOY, Inc. Consumer Class Action Litig.*, No. 14-cv-00428-MMM-JEMx, 2015 WL 12732461 (C.D. Cal. May 27, 2015).

historical quantity of product in the but-for world, in which some consumers would potentially have lower willingness to pay for the but-for product. From an economic perspective, the assumption of a constant or “fixed” quantity supplied results in an estimate of damages based solely on a downward shift in demand that the plaintiffs contend would occur in the but-for world, measured at the quantity sold in the actual world.⁹⁰ Other experts have assumed that the quantity sold in the but-for world may be different from the quantity sold in the actual world.⁹¹ For example, some experts have estimated a but-for price based on a market simulation in which supply-side factors include hypothetical sellers that, along with the defendant, maximize their individual profits taking into account their competitors’ reactions.⁹² The validity of the estimates from this method depends on how accurately the assumptions and characteristics of the market simulation reflect the characteristics of the actual market.

Estimation of Economic Damages in Securities Cases

Consider the typical securities case,⁹³ with a plaintiff investor (often, on behalf of a class of investors) who claims to have suffered damages from misrepresentations (false or misleading statements or omissions) made by a defendant (or group of defendants) in connection with the purchase of a company’s stock.⁹⁴ In such cases, the plaintiff claims to have suffered investment losses upon the revelation to the market of the false or misleading nature of the misrepresentations through one or more “corrective” disclosures or events. This section introduces key economic issues and areas of disagreement among experts in the analysis of damages in such cases.⁹⁵

90. For this reason, the but-for price calculated using this assumption may not be a market price that would result from the interaction of supply and demand in the relevant market.

91. See, e.g., *Earl v. Boeing Co.*, 611 F. Supp. 3d 345 (E.D. Tex. 2020).

92. See, e.g., Greg M. Allenby et al., *Valuation of Patented Product Features*, 57 J.L. & Econ. 629, 630 (2014), <https://doi.org/10.1086/677071>.

93. Statutes under which plaintiffs in securities cases may seek to recover damages include, but are not limited to, the Securities Exchange Act of 1934 § 10(b) (and Rule 10b-5 promulgated thereunder) (“Rule 10b-5” cases) and the Securities Act of 1933 § 11 or § 12(a)(2) (“Section 11” or “Section 12” cases).

94. The concepts discussed herein also apply in securities cases where securities other than common stock (e.g., options and bonds) are at issue.

95. As a threshold matter, given the nature of the plaintiff’s claims, analysis of damages in securities cases requires that experts have extensive training in financial economics, the discipline within economics that studies security prices and their determinants. Such training is often obtained through doctoral degrees in economics, finance, or accounting, followed by additional experience (in an academic or professional setting) in the analysis of the determinants of security prices and values. Depending on the circumstances, additional training may further bolster the

Inflation and Losses Caused

The economic issues that arise in the analysis of damages in securities cases relate to the concepts of inflation and losses caused. This section introduces these two concepts and how they relate to the measurement of damages in securities cases.

The concept of stock price inflation arises, for example, when analyzing damages in Rule 10b-5 cases, where the “out-of-pocket” measure of damages (calculated as the difference between inflation at the time of purchase and inflation at the time of sale)⁹⁶ limits the amount of per-share damages that plaintiffs can recover.⁹⁷ Inflation at a given point in time is the difference between (1) the actual price at which the stock then traded and (2) the but-for price that would have existed in the absence of the misrepresentations (also known as the “true” value).⁹⁸

The actual stock price is easily ascertained based on the plaintiff’s trading records or market data. However, estimation of the but-for price—and thus, inflation—requires two analytical steps. The first step is a careful delineation of the information that the plaintiff claims the defendants could and should have revealed instead of or in addition to the misrepresentations prior to the plaintiff’s purchase (the “but-for disclosure” or “truthful substitute”). This delineation is based on the plaintiff’s allegations and typically not on economic analysis. However, this delineation is critical because the second step, which is based on economic analysis, is the application of tools and techniques from financial economics to estimate the stock’s but-for price given the but-for disclosure.

The concept of losses caused arises, for example, when analyzing damages in Rule 10b-5, Section 11, and Section 12(a)(2) cases, where the plaintiff’s recoverable

expert’s qualification to analyze issues at hand (e.g., accounting training may be helpful in securities cases where the misrepresentations relate to a company’s accounting disclosures and practices).

This section focuses on the analysis of damages on a per-share basis. Total damages for a plaintiff are calculated by combining damages per share and the plaintiff’s trading records (with the number of shares bought and sold by the plaintiff on relevant dates). In securities class actions, total damages are typically not covered in expert reports submitted to the court, as class members’ trading records only become available after damages claims are filed, typically after a settlement occurs or if a verdict for plaintiffs is reached.

96. See, e.g., *Robbins v. Koger Props., Inc.*, 116 F.3d 1441, 1447 n.5 (11th Cir. 1997).

97. Per-share recoverable damages in Rule 10b-5 cases are the lesser of (1) “out-of-pocket” damages, (2) the losses caused by the alleged misrepresentations (discussed further below), and (3) a cap prescribed by the Private Securities Litigation Reform Act of 1995 (the “90-day bounce-back” provision).

98. See, e.g., *Ludlow v. BP, P.L.C.*, 800 F.3d 674, 682 (5th Cir. 2015) (“At the same time, the ‘out-of-pocket measure,’ sometimes called the ‘price inflation’ metric, is often used. Under this theory, ‘a purchaser of securities may recover against a defendant . . . only the “difference between the price paid and the [true] value of the security . . . at the time of the initial purchase by the defrauded buyer.”’”).

damages are limited to the losses, if any, caused by the misrepresentations.⁹⁹ In the context of damages, losses caused refers to stock price declines, if any, that occurred in the actual world and are attributable to the arrival of information (corrective information) into the market that is causally linked to the misrepresentations.

Quantification of the losses, if any, caused by the misrepresentations requires, as a first step, identification of instances when new corrective information was revealed to the market. The next step is to use tools and techniques from financial economics to measure, for each of those instances, the impact, if any, of such information on stock price. Such measurement is typically conducted using tools and techniques from financial economics. For example, an event study model, discussed below, can be used to estimate the impact that market or industry factors—which are typically not causally linked to the misrepresentations—had on the stock price. In addition, any effect of company-specific information that is not corrective information (confounding information) must be separated from any effect of company-specific information that is. In sum, any losses that resulted from “changed economic circumstances, changed investor expectations, new industry-specific or firm-specific facts, conditions, or other events,”¹⁰⁰ and thus would have occurred even in the absence of corrective information, must be isolated and removed in determining the amount of any stock price declines that are attributable to the misrepresentations.

99. In the context of Rule 10b-5 cases, in *Dura Pharms., Inc., v. Broudo*, 544 U.S. 336, 342–44 (2005), the Supreme Court referred to losses caused as a limitation on out-of-pocket damages: “If the purchaser sells later after the truth makes its way into the marketplace, an initially inflated purchase price might mean a later loss. But that is far from inevitably so. When the purchaser subsequently resells such shares, even at a lower price, that lower price may reflect, not the earlier misrepresentation, but changed economic circumstances, changed investor expectations, new industry-specific or firm-specific facts, conditions, or other events, which taken separately or together account for some or all of that lower price. . . . Indeed, the Restatement of Torts, in setting forth the judicial consensus, says that person who ‘misrepresents the financial condition of a corporation in order to sell its stock’ becomes liable to a relying purchaser ‘for the loss’ the purchaser sustains ‘when the facts . . . become generally known’ and ‘as a result’ share value ‘depreciate[s].’” In the context of Section 11 and Section 12(a)(2), defendants can remove from the damages amount prescribed by statute (*statutory damages*) the portion that did not result from the misrepresentations. See 15 U.S.C. § 77(k) regarding Section 11 (“[I]f the defendant proves that any portion or all of such [statutory damages] represents other than the depreciation in value of such security resulting from [the alleged material misrepresentations], such portion of or all such damages shall not be recoverable”) and 15 U.S.C. § 77l regarding Section 12 (“if the [defendants prove] that any portion or all of [statutory damages] represents other than the depreciation in value of the subject security resulting from [the alleged material misrepresentations], then such portion or amount . . . shall not be recoverable.”).

100. *Dura*, 544 U.S. at 343.

Key Areas of Disagreement in the Economic Analysis of Inflation and Losses Caused

Event study model

Damages analyses in securities cases almost invariably involve an event study model, which is a widely used and generally accepted methodology to analyze changes in stock prices following the revelation of information to the market.¹⁰¹ Using an event study model to analyze changes in stock price following the release of allegedly corrective information is a common step in the analysis of losses caused.¹⁰²

An event study model can be used to estimate the portion of a stock's return (the percentage change in stock price, including dividends) over some period (the event window) that can be attributed to market and industry factors. The remaining portion, often labeled the residual or abnormal return, may be attributed to company-specific information (i.e., information that affects the company but not the overall market or its industry peers),¹⁰³ including any corrective information disclosed over the event window. An event study model can be used to test whether a residual return is statistically unusual relative to the typical level of stock-price volatility (i.e., "statistically significant").¹⁰⁴ Economists interpret residual returns that are statistically significant as evidence that company-specific information caused a stock-price change. Residual returns that are not statistically significant are interpreted as having no cause that can be distinguished from randomness.

There are several common areas of disagreement among experts when using an event study model. Each of these areas illustrates the importance of carefully

101. See, e.g., A. Craig MacKinlay, *Event Studies in Economics and Finance*, 35 J. Econ. Literature 13 (1997), <http://www.jstor.org/stable/2729691>.

102. An event study model can also be used to analyze changes in stock price following allegedly false or misleading statements, which may be relevant for the analysis of inflation.

103. Note that company-specific information does not necessarily coincide with information disclosed by the company. Sometimes, company disclosures may convey information that moves the overall market and the stock prices of its industry peers. See, e.g., George Foster, *Intra-Industry Information Transfers Associated with Earnings Releases*, 3 J. Acct. & Econ. 201 (1981), [https://doi.org/10.1016/0165-4101\(81\)90003-3](https://doi.org/10.1016/0165-4101(81)90003-3); Andrew J. Patton & Michela Verardo, *Does Beta Move with News? Firm-Specific Information Flows and Learning About Profitability*, 25 Rev. Fin. Stud. 2789–2839 (2012), <https://doi.org/10.1093/rfs/hhs073>. And sometimes, third-party disclosures may convey company-specific information (e.g., a short-seller report analyzing a company's prospects, as discussed in Alexander Ljungqvist & Wenlan Qian, *How Constraining Are Limits to Arbitrage?*, 29 Rev. Fin. Stud. 1975, 2012–14 (2016), <https://doi.org/10.1093/rfs/hhw028>).

104. The most commonly used threshold for the assessment of statistical significance in event studies is the 5% level. As discussed in David H. Kaye & Hal S. Stern, *Reference Guide on Statistics and Research Methods*, in this manual, "In practice, statistical analysts typically use levels of 5% and 1%. The 5% level is the most common in social science, and an analyst who speaks of significant results without specifying the threshold probably is using this figure."

tailoring the event study model in a particular case to the facts and circumstances of that case. Absent such tailoring, experts may reach erroneous conclusions.

First, experts may disagree on whether the event study model adequately captures the effect of market and industry factors. There may be disagreement about whether appropriate market and industry factors¹⁰⁵ have been selected,¹⁰⁶ or about whether the model adequately captures the sensitivity of company stock returns to market and industry factors.¹⁰⁷ A model that overstates or understates the effect of market or industry factors will incorrectly estimate the residual return and thus may lead to an incorrect finding of statistical significance (or lack thereof).¹⁰⁸

Second, experts may disagree on whether a model adequately estimates the typical level of stock-price volatility used to assess the statistical significance of residual returns.¹⁰⁹ The typical level of volatility may vary significantly over time for a given company stock, as circumstances change.¹¹⁰ A model that overstates or understates volatility may lead to an incorrect finding of statistical significance (or lack thereof).¹¹¹

Third, experts may disagree on the appropriate event window to analyze residual returns following a particular information release. In an efficient market, a company's stock price will at all times quickly and fully reflect all publicly

105. Market and industry factors are usually modeled using the returns on market-wide and industry indices, respectively. In choosing an appropriate industry index, experts typically balance between an index with more constituent company stocks (less prone to the effect of outlier returns but also potentially less comparable to the stock of interest) and an index with fewer constituents (more comparable to the stock of interest but also more prone to the effect of outlier returns).

106. See, e.g., Exhibit 138 to Defendants' Motion for Summary Judgment, *Strathclyde Pension Fund v. Bank OZK*, No. 4:18-cv-793-DPM, ECF No. 158–9, (E. D. Ark. Feb. 5, 2022).

107. See, for example, the discussion of the dispute about whether a model adequately captured the sensitivity to industry factors in *Smilovits v. First Solar Inc.*, No. CV12-0555-PHX-DGC, 2019 WL 7282026, at 11–12 (D. Ariz. Dec. 27, 2019).

Academic literature in financial economics shows that the sensitivity of company stock returns to market and industry factors varies depending on the circumstances. See, e.g., Patton & Verardo, *supra* note 103, at 2789.

108. The statistical significance of residual returns is assessed using a t-statistic, which is calculated as the ratio of the residual return to the estimate of typical volatility. Thus a misestimated residual return will lead to a misestimated t-statistic.

109. See, e.g., Defendants' Motion for Partial Summary Judgment, *Evanston Police Pension Fund v. McKesson Corp.*, No. 3:18-cv-6525-CRB, ECF No. 166, (N.D. Cal. June 7, 2021).

110. See, e.g., Robert Engle, *GARCH 101: The Use of ARCH/GARCH Models in Applied Econometrics*, 15 J. Econ. Persps. 157, 158 (2001), <https://doi.org/10.1257/jep.15.4.157> (“Even a cursory look at financial data suggests that some time periods are riskier than others; that is, the expected value of the magnitude of error terms at some times is greater than at others.”).

111. Misestimated volatility will cause a misestimated t-statistic, which is used to assess statistical significance, as discussed *supra* note 108.

available information.¹¹² Thus, if the stock trades in an efficient market, experts will typically analyze stock-price changes over short event windows (such as the trading day that immediately follows the release). However, under certain circumstances, experts may choose to apply different event windows. For example, in the presence of multiple pieces of information disclosed at different times during the same trading day, event windows shorter than a trading day may be helpful in analyzing the effect of a particular piece of information.

Analysis of “confounding” information

Because an event study model estimates a single residual return for a particular event window, it does not disaggregate the impact of separate pieces of company-specific information arriving within the same event window. This feature of event study models limits its usefulness in securities cases when the arrival of corrective information coincides with the arrival of confounding information.

To value the effect of corrective information when there is confounding information, experts may need to apply available tools and techniques from financial economics (other than an event study model) that can be used to value specific pieces of information (e.g., discounted cash-flow models, valuation based on multiples, earnings response models).¹¹³ To evaluate whether a specific application of such tools and techniques is economically sound, it is necessary to scrutinize the underlying assumptions and determine if they are sufficiently tethered to the allegations, facts, and circumstances of the case. Without such scrutiny, it is not possible to determine a priori whether a particular method or group of methods will be capable of estimating losses caused by misrepresentations consistent with the plaintiff’s theory of liability in the case.

Measuring inflation at the time of purchase

Because inflation at each point in time (e.g., at the time of the plaintiff’s purchase) cannot be observed from market prices (indeed, it is based on a counterfactual, or

112. This is known as the semi-strong form of market efficiency. See Stephen A. Ross, Randolph Westerfield & Bradford D. Jordan, *Corporate Finance* 346 (6th ed. 2002) (“[S]emistrong-form efficiency requires not only that the market be efficient with respect to historical price information, but that all of the information available to the public be reflected in price.”).

113. See, e.g., Richard A. Brealey, Stewart C. Myers & Franklin Allen, *Principles of Corporate Finance* (11th ed. 2014).

The concept of market efficiency, discussed above, is relevant to the analysis of information that may have impacted the stock price. In an efficient market, only new information can impact the stock price and no price declines can be attributed to the repetition of information that is already publicly available.

but-for, disclosure), it must be estimated. Experts often disagree on how to measure inflation, including whether stock-price declines following the revelation of corrective information can be used to measure inflation earlier in time.

Stock-price declines following corrective information can measure inflation earlier in time only if (1) any price impact of confounding information has been removed, as discussed above, (2) the corrective information matches the but-for disclosure, and (3) the information environment (i.e., market, industry, and company conditions) earlier in time is such that if the company had made the but-for disclosure then, the stock price would have declined by the same amount as it did upon the release of corrective information in the actual world.¹¹⁴ In its *Goldman Sachs* ruling, the Supreme Court noted that the “inference—that the back-end price drop equals front-end inflation—starts to break down when there is a mismatch between the contents of the misrepresentation and the corrective disclosure.”¹¹⁵ If the facts and circumstances of the case do not support using stock-price declines following corrective information to measure inflation earlier in time, then additional analysis, including analysis based on valuation tools and techniques from financial economics, is typically required.¹¹⁶

Estimation of Economic Damages in Antitrust Cases

This section provides examples to illustrate the application of the principles for estimation of economic damages to antitrust cases. The scope of the discussion is limited; a full discussion of damages methods across the complete spectrum of antitrust offenses is beyond the scope of this reference guide.¹¹⁷

Threshold Issues

A model for economic damages arising from an antitrust violation must satisfy the causation requirement articulated in *Comcast*, by measuring the economic damages arising only from defendant conduct found to be unlawful.¹¹⁸

114. See, e.g., Bradford Cornell & R. Gregory Morgan, *Using Finance Theory to Measure Damages in Fraud on the Market Cases*, 37 UCLA L. Rev. 883, 889–97 (1990).

115. *Goldman Sachs Grp., Inc. v. Arkansas Tch. Ret. Sys.*, 594 U.S. 113, 123 (2021).

116. As is the case when valuing the effect of corrective information when there is confounding information, there is no a priori reason that a particular method or particular group of methods will be capable of estimating inflation consistent with the plaintiff’s theory of liability in a specific securities case.

117. For a comprehensive discussion of antitrust damages, see A.B.A., *Proving Antitrust Damages: Legal and Economic Issues* (2017), and Daniel L. Rubinfeld, *Antitrust Damages*, in *Research Handbook on the Economics of Antitrust Law* 378–93 (Einer Elhague ed., 2012).

118. In *Comcast*, the court held that the plaintiffs failed to present a valid methodology for measurement of class-wide damages because their damages model included the effects of conduct

Further, a model estimating damages arising from an antitrust violation must measure what courts have described as antitrust injury. Plaintiffs must show not only that they suffered economic injury, but additionally that the injury they sustained is “of the type the antitrust laws were intended to prevent and that flows from that which makes defendants’ acts unlawful.”¹¹⁹

In many instances, meeting this requirement can be straightforward. For example, any alleged overcharge that consumers pay due to a price-fixing cartel flows directly from the cartel’s subversion of price competition, which the antitrust laws seek to prevent. Hence, damages based on a reliable overcharge computation reflect antitrust injury.

However, not all harms constitute antitrust injury. For example, a firm may be harmed by two rivals that merge to form a single, more efficient competitor, but the harm that firm sustains—an erosion in its market position because it faces a new, more efficient competitor—is the result of increased competition rather than harm to competition.¹²⁰ An increase in competition is not something the antitrust laws seek to prevent—quite the opposite, in fact—and as such it is not antitrust injury.¹²¹

Quantifying Damages in an Antitrust Matter

The foundation of an antitrust damages model is a but-for world that reflects all relevant conditions in the actual world other than the anticompetitive conduct at issue, as explained in the section titled “The Standard General Approach to Quantification of Economic Damages” above.

Models used to compute antitrust damages generally fall into one of two categories, overcharge damages and lost-profit damages. As explained below, the choice of model depends on the nature of the plaintiff’s claim and the economic relationship between the plaintiff and defendant.

Overcharge Damages

An overcharge model attempts to measure damages sustained by one of the parties to a vertical buyer–seller relationship by deriving an estimate of what the price in the market would be if the conduct did not occur.

under multiple theories of liability, only one of which was found to be triable on a class-wide basis. See *Comcast Corp. v. Behrend*, 569 U.S. 27 (2013).

119. See *Brunswick Corp. v. Pueblo Bowl-O-Mat, Inc.*, 429 U.S. 477 (1977).

120. *Cargill, Inc. v. Monfort of Colorado, Inc.*, 479 U.S. 104 (1986).

121. See A.B.A., *supra* note 117, at 19–33. See also Jonathan M. Jacobson & Tracy Greer, *Twenty-One Years of Antitrust Injury: Down the Alley with Brunswick v. Pueblo Bowl-O-Mat*, 66 *Antitrust L.J.* 273 (1998).

For example, a plaintiff that purchased a product from a supplier engaged in anticompetitive conduct might compute overcharge damages as the difference between the supra-competitive price actually paid and the lower price that would have prevailed absent the conduct at issue (i.e., the but-for price), multiplied by the quantity actually purchased. Similarly, a supplier that sold a product to a buyer engaged in anticompetitive conduct (e.g., a monopsonistic conspiracy) may compute overcharge damages as the difference between the higher price it would have received absent the buyer's anticompetitive conduct and the artificially suppressed price actually received, multiplied by the quantity actually sold. Antitrust claims that lend themselves to an overcharge model of damages include horizontal price-fixing, bid-rigging, illegal-monopolization, and exclusionary-conduct claims when brought by a customer or supplier.

Damages experts generally rely on one of two approaches to estimate overcharge damages:

1. A before-and-after model, which compares price during the period in which the anticompetitive conduct is believed to have had an effect (the impact period) with price in a period before (or after) the anticompetitive conduct occurred (the control period).
2. A difference-in-differences model, which compares the change in price from the control period to the impact period with the change in price during the same periods in a reasonably comparable market.

Conceptually, both methods seek to identify price levels absent the impact of the challenged conduct and then estimate the competitive price level in the but-for world. Below is a hypothetical example of a price-fixing cartel to illustrate the application of these methods.

Before-and-after model of overcharge damages

Suppose that the plaintiffs allege harm due to supra-competitive prices charged by a cartel and estimate damages as follows: (1) estimate a regression model with observations on prices over a period of months or years prior to the date at which the cartel took effect,¹²² under the assumption that the model describes prices that would have prevailed but for the cartel; (2) use the estimated model to

122. Multiple regression is a statistical method that can be used to measure the relationship between multiple explanatory variables and a variable of interest (e.g., price), and thus distinguish between the effect of defendant's anticompetitive conduct and the effect of other economic factors. A detailed discussion of the use of regression analysis (including to estimate economic damages) is beyond the scope of this reference guide. See Daniel L. Rubinfeld & David Card, *Reference Guide on Multiple Regression and Advanced Statistical Models*, in this manual. See also A.B.A., *supra* note 117, at 123.

predict the prices that would have prevailed during the damages period but for the cartel; and (3) estimate overcharge damages as the difference between observed prices during the damages period and prices predicted by the regression model, multiplied by the quantity sold.¹²³

The validity of such an estimate depends on the validity of the underlying regression model, which must account for the impact on prices of economic factors, other than the challenged conduct, that influence demand and costs. A regression that omits such factors will generally yield an unreliable estimate of damages. The direction of the error depends on the nature of the omitted factors. Two examples illustrate the point:

Example: Negative news during the cartel period about the health effects of the product at issue causes demand to fall, and absent the cartel, the price of the product at issue would fall as a result. A regression model that fails to account for the price effect of the negative news would likely overstate but-for prices during the cartel period and understate damages as a result.

Example: The structure of the market changed due to a merger between rival manufacturers just before the cartel began operating. The price effects of the merger, if any, would persist in the but-for world, as the cartel did not cause the merger. A regression model that failed to separately account for the price effects of the merger would improperly attribute the effects of the merger to the cartel and overstate damages as a result.

A regression model that supports a before-and-after estimate of damages may also yield an unreliable estimate of but-for prices if the conduct at issue also affects the explanatory factors included in the regression. Examining the impact of the conduct at issue on explanatory factors in a regression used to estimate but-for prices is thus a critical step in evaluating a benchmark model of damages.

123. Note that this approach assumes that the impact of explanatory variables is identical during the control and cartel periods. Rather than forecasting but-for prices in the cartel period based on an earlier control period, a damages expert may instead rely on the so-called dummy variable method: (1) estimate a regression model with observations on price using a sample that includes the damages period in addition to the period beforehand; (2) include in the regression specification a dummy variable that measures the difference between prices in the control and cartel periods, holding all other factors constant; and (3) use the point estimate for the dummy variable as the basis for computing overcharge damages. The dummy variable method allows a more flexible specification in which the effects of explanatory variables can differ during the control and cartel periods. See Justin McCrary & Daniel L. Rubinfeld, *Measuring Benchmark Damages in Antitrust Litigation*, 3 J. Econ. Methods, no. 1, 2013, at 63–74, <https://doi.org/10.1515/jem-2013-0006>. See also John K. Wald, *Antitrust Damages in Financial Markets*, 16 J. Competition L. & Econ., no. 1, 2020, at 63–73, <https://doi.org/10.1093/joclec/nhaa003>, for an example of the application of this method in the NASDAQ odd-eighths litigation.

Example: Suppose that the price of the product at issue depends on the price of a particular input, and the regression model includes the price of that input. If cartel participants account for a substantial fraction of demand for the input, then in the but-for world, the price of the input will likely be lower.¹²⁴ Thus a regression that relies on the actual input price, which reflects the impact of the cartel, will likely yield a biased estimate of the but-for price of the product at issue.¹²⁵

Difference-in-differences model of overcharge damages

The preceding examples highlight the errors that may arise if a regression-based estimate of overcharge damages omits economic factors that affect price. The omitted factors discussed above—new information and changes in market structure—are observable. In such cases, correcting the bias may be as simple as including appropriate explanatory variables in the regression to measure the impact of the omitted factors on price.

Many cases, however, are not so simple. The omitted factors that affect prices may be unobservable, a variable that can serve as a proxy for unobserved factors may not be available, or the nature of the omitted factor is such that its impact cannot be separately identified using multiple regression. In such cases, damages experts may turn to a difference-in-differences regression model in order to account for the impact of omitted factors.

Rather than estimating the overcharge by examining price before and during the cartel period, a difference-in-differences model isolates the impact of the cartel through comparison to a second, comparable market unaffected by the conduct at issue. A difference-in-differences regression measures the impact of the cartel as the difference between (1) the change in prices in the market of interest, and (2) the change in prices over the same period in a second, comparable market unaffected by the conduct at issue.¹²⁶ The key assumption here that identifies the impact of the cartel—the so-called parallel trend assumption—is that prices in the two markets share a common, parallel

124. The cartel suppresses output of the final good and thus suppresses demand for the input in question. If cartel participants account for a substantial fraction of demand for the input, increased demand for the input in the but-for world means that the price of the input will likely increase as well.

125. For a more detailed discussion of this point, see Halbert White, Robert Marshall & Pauline Kennedy, *The Measurement of Economic Damages in Antitrust Civil Litigation*, 6 A.B.A. Econ. Comm. Newsl., no. 1, 2006, at 18.

126. See, e.g., Joshua D. Angrist & Jörn-Steffen Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion* (2008).

trend, so that absent the cartel, change in prices in the two markets over time would be identical.¹²⁷

For example, suppose the plaintiffs claim that suppliers of a product in a particular geographic region operated a cartel. A damages expert might adopt as a comparator a second, separate geographic market for the same product in which the alleged conspirators do not operate, and use a difference-in-differences regression to account for the effects of unobserved factors and isolate the price effects of the cartel.¹²⁸

Lost-Profits Damages

A lost-profits model is frequently used to measure damages in matters in which the plaintiff and defendant are competitors rather than parties to a vertical, buyer–seller relationship, such as matters involving exclusionary conduct.

While overcharge damages require an estimate of but-for prices, a damages model that seeks to measure lost profits requires a more extensive counterfactual that specifies prices, unit sales (or market share), and costs in the but-for world.¹²⁹ Assuming that the plaintiff’s fixed costs would not change in the but-for world, lost-profit damages for a firm harmed by a rival’s anticompetitive conduct are given by the following formula¹³⁰:

$$D = (P_{\text{but-for}} - C_{\text{but-for}}) \times Q_{\text{but-for}} - (P_{\text{actual}} - C_{\text{actual}}) \times Q_{\text{actual}}$$

A model that estimates lost profits must carefully consider the effects of the conduct at issue on each element of the plaintiff’s profits in the but-for world.

Price Effects. Because conduct that partially or completely excludes a firm from a market reduces competition and allows the defendant to increase prices, prices in the but-for world will likely be lower than observed prices. Indeed, the prospect of increasing prices creates an incentive for firms to engage in exclusionary conduct.¹³¹ That said, in some circumstances plaintiffs may assert that in the but-for world, they would be able to charge higher prices, as the

127. Note that the regression specification should still include all observable factors that affect price and are independent of the cartel (i.e., do not vary as a result of the cartel), just as with the before-and-after model discussed above.

128. See, e.g., R. Forrest McCluer & Martha A. Starr, *Using Difference in Differences to Estimate Damages in Healthcare Antitrust: A Case Study of Marshfield Clinic*, 20 Int’l J. Econ. Bus. 447 (2013), <https://doi.org/10.1080/13571516.2013.800323>.

129. Some damages experts may specify damages using a model of but-for profit margins rather than modeling underlying prices and costs.

130. Costs shown here should be read as average cost per unit of output, not as variable cost.

131. Note, however, that some theories of liability, such as predatory pricing, imply higher prices in the but-for world.

challenged conduct prevented them from effectively competing and forced them to offer prices lower than those they otherwise would have offered.¹³²

Quantity Effects. Economic theory teaches that if prices increase, output declines. Thus, if the effect of the conduct at issue is to raise prices, lower prices in the but-for world will generally be accompanied by higher output. However, the critical question is not output for the market as a whole but whether the plaintiff's output would increase absent the challenged conduct.

Lower prices and higher output have offsetting effects on profits: All else being equal, lower prices reduce profits, but higher output increases profits. The net impact of these effects is an empirical question whose answer depends on the elasticity of demand for the products or services at issue. If demand is elastic, quantity effects will outweigh price effects, profits in the but-for world will increase, and all else being equal, the plaintiff's profits in the but-for world would be higher. If demand is inelastic, profits in the but-for world will decrease, and all else being equal, the plaintiff's profits in the but-for world would be lower.

Cost Effects. The impact of the challenged conduct on the plaintiff's costs is likewise an empirical question; the plaintiff's costs can increase or decrease as a result of the challenged conduct. In some cases, the conduct at issue may raise the plaintiff's costs. Indeed, the specific objective of exclusionary conduct is often raising a rival firm's costs. Additionally, rather than directly raising input costs, the conduct at issue may increase the plaintiff's costs by reducing the plaintiff's output and thereby depriving the plaintiff of the benefit of economies of scale. In such cases, the plaintiff's unit costs would decline as output increases in the but-for world.

On the other hand, the plaintiff's costs in the but-for world may be higher if the output increase contemplated in the damages model exceeds the plaintiff's actual capacity to produce, market, or distribute the products or services at issue and would require additional investment. In such cases, the plaintiff must establish that it had the wherewithal to increase its capacity in the but-for world, and estimated damages should reflect the costs of any incremental investments necessary to expand the plaintiff's capacity.

Models of lost profits frequently rely on the assumption that the period prior to the conduct at issue provides a reliable basis for predicting market conditions in the but-for world. As explained in the section titled "Considering All the Differences in the Plaintiff's Situation in the But-For Scenario Versus Assuming That Many Aspects Would Be the Same as in Actuality" above, this assumption requires careful scrutiny, as it may yield an unreliable estimate of damages. The purpose of the defendant's challenged conduct may have been to forestall

132. See, e.g., *ZF Meritor, LLC v. Eaton Corp.*, 696 F.3d 254 (3d Cir. 2012). ZF Meritor's lost-profits damages model relied on projections that assumed that the exclusionary loyalty discounts offered by defendant Eaton forced ZF Meritor to offer deep discounts that it otherwise would not have offered.

changes in competitive conditions (e.g., by deterring entry) and maintain the status quo, in which case market conditions in the period prior to the challenged conduct are not a reliable basis for predicting market conditions in the but-for world. Suppose, for example, that the market for a particular product was a duopoly, and that one firm (the defendant) entered an exclusive dealing agreement with a downstream purchaser that foreclosed the second firm (the plaintiff) and deterred the entry of new competitors. Then a damages model that assumes a duopoly in the but-for world would yield an unreliable conclusion, as it fails to account for the competitive impact of nascent competitors that would have entered the market but for the challenged conduct.

Estimation of Economic Damages from Loss of Personal Income

Many of the disputes that arise in estimating damages for lost personal income can be resolved by carefully applying the standard damages approach. Damages are the difference between the but-for and actual worlds, where the actual world reflects any mitigating factors. Estimating such damages can also involve issues that are unique, such as estimating losses over a person's lifetime, valuing fringe benefits, estimating lost income in wrongful death cases, or estimating compensation for shortened life expectancy. We discuss these issues below.

Estimation of Losses Over a Person's Lifetime

In nearly all cases involving lost income, the effects continue past trial and sometimes until the plaintiff's death. Therefore, quantifying damages for loss of personal income necessarily involves projecting the plaintiff's work history and retirement. Conceptually, the estimate of income for each year, either but-for or actual, is the expected income multiplied by the probability that the person will be working for that year. The probability that the person will be working for that year is the product of the probability that the person will survive the year, the probability that the person will be in the labor force, and the probability that the person will be employed, given that the person worked in the prior year.¹³³ We refer to this as the standard method for estimating personal losses.

In many cases, such as those involving wrongful termination, the projection that the plaintiff was working is the same for both the but-for world and the actual

133. Except for wrongful-death cases, the probability that the persons will be working is usually 1 for both the but-for and actual cases. In wrongful-death cases, the probability is still usually 1 in the but-for case but 0 in the actual case.

world. However, in wrongful death cases and some personal injury cases, these projections may differ,¹³⁴ and the expert will need to compute separate projections for the but-for and actual worlds before taking the difference between the two.

These projections may rely on data from government agencies, including the Bureau of Labor Statistics (BLS) and the Social Security Administration (SSA).¹³⁵ These data include tables on survival, workforce participation, and employment. The expert usually needs to manipulate this information in order to generate the conditional probabilities needed.

In order to simplify the calculations, the expert may use the person's expected lifespan and retirement age based on the person's age at trial using standard tables from the SSA and BLS. Then the expert need only sum the discounted losses for each year until the expected age at death. This method, often referred to as the life-expectancy method, will considerably simplify the calculations associated with estimating lost retirement benefits. However, the standard and the life-expectancy methods will usually generate the same estimate of losses only if the expert is assuming that the expected discounted income in each year is the same—that is, that the expected increase in income is offset by the discount rate (see section titled “Disagreements About the Discount Rate to Value the Impact of the Alleged Harmful Act in Periods After the Valuation Date” above).

Calculation of Fringe Benefits

Fringe benefits are often a component of lost pay and may include medical insurance and retirement benefits such as Social Security. Although sick days and vacation are also fringe benefits, they are included already in lost-pay calculations. An exception occurs if these days are accrued but not taken, where, for example, the employee lost the cash payout that would have occurred at a normal termination or lost the benefit of future days off with pay.

Medical insurance benefits

In the following discussion, we assume that the plaintiff no longer has the benefit of the employer-provided insurance as a result of the actions of the defendant. Such situations typically arise in wrongful-termination or wrongful-death cases.

134. This situation may arise in a personal injury case if the injured person is less likely to be able to work as a result of the accident. If so, then the person may have an increased likelihood of leaving the labor force for each age compared to the likelihood prior to the incident. Similarly, the injured person may be more likely to retire at an earlier age.

135. Data from additional providers may be available with a finer level of granularity than SSA or BLS data (e.g., with respect to location or occupation). However, the reliability of these data may be disputed because of questions about the methodology used to generate the tables.

Reference Guide on Estimation of Economic Damages

Calculating damages for lost medical insurance is straightforward if the plaintiff can purchase insurance under COBRA, or from their current employer, or on the open market. Then the value of the lost medical insurance is the employee's portion of the premium. If insurance coverage available to the plaintiff differs significantly between the but-for and the actual worlds, then the expert will need to project the impact of the difference in policy coverage.

If the plaintiff chooses not to purchase insurance even though the option is available, or the plaintiff is unable to purchase insurance,¹³⁶ then the plaintiff may argue that the value of insurance is the sum of their actual expenses less the premium in the but-for world. The defendant will likely respond that the plaintiff assumed the risk that the incurred medical expenses could exceed the plaintiff's portion of the premium, and therefore that the defendant's responsibility should be limited to the plaintiff's portion of the premium forgone. The plaintiff may counter that their pay was insufficient to afford the insurance.

Retirement benefits

For lost retirement benefits, the issues are similar to those involving lost medical benefits, but the calculations are more complex. There are basically two types of retirement plans: defined benefit plans and defined contribution plans. Defined benefit plans are those where the benefits paid out after retirement are guaranteed to be a definite amount upon retirement. In contrast, defined contribution plans are those where the employer makes a predefined contribution for the employee but the benefits paid out depend on the return earned on the money invested.

The expert can calculate both types of retirement plans on the basis of either the amounts paid in or the amounts paid out by the employer. If the expert uses the amount the employer paid for the benefits, which may be a function of the amount the plaintiff earned, then the calculation is analogous to computing the loss in plaintiff's earnings. The disadvantage of this approach is that the amounts paid in may not adequately predict the benefits paid out, particularly when the plan is a defined benefit plan. We discuss this topic below in connection with Social Security benefits, where the problem is particularly acute.

Defined benefit plan

To determine the present value of the benefits received under a defined benefit plan, the calculation is simplified if the expert uses the life-expectancy method

136. For example, a preexisting condition may make it difficult to purchase insurance on the open market or may limit the plaintiff's coverage to exclude a preexisting condition either permanently or for a period of time.

to calculate the plaintiff's losses. In this situation, the expert must determine the number of years that the plaintiff would have worked at the firm upon retirement, and the plaintiff's retirement age, expected lifespan, and salary at the firm over time. These factors must be consistent with the expert's assessment of the projected trajectory of the plaintiff's employment in both the but-for and the actual worlds.

However, if the expert instead uses probability tables for each year, then the calculation is more complex. For each year where the plaintiff may cease to be in the labor force for reasons other than death, the expert must determine the likelihood that the plaintiff is receiving benefits from the plan because of disability (if the plan permits) or retirement. Complicating this determination is that the payout from the plan may depend on the age at retirement. Thus, the calculation must incorporate the probability for each possible payout. Depending on the plan, defining possible outcomes can be extremely complex.

A common example of a defined benefit plan is Social Security.¹³⁷ Determining benefits from Social Security can be forbiddingly complex because the number of potential outcomes is so large. For example, a person can retire at almost any age, and disability payments are made if the person is unable to work. In addition, calculating the benefit at any age depends on the person's average salary over the most recent thirty-five years. If Social Security benefits are critical to the magnitude of damages, the expert may choose to simplify the calculation by relying on the life-expectancy method.

Defined contribution plan

For a defined contribution plan, the expert's task is to project the employer's contribution, the number of years that the employee would have worked at the firm, and the employee's age at retirement. Generally, this determination is straightforward because it is based on the same factors the expert would use to project the employee's salary in the but-for and actual worlds. The present value of the employer's contributions will be the expected payouts from the plan.

Wrongful Death

Generally, calculation of economic damages for wrongful death depends on whether the claimant is a relative of the decedent or is the estate. If the claimant is a relative of the decedent, economic damages are limited to the economic value that the relative would have received had the decedent lived. If the relative

137. Social Security is generally regarded as a defined benefit plan although it has some elements of a defined contribution plan.

Reference Guide on Estimation of Economic Damages

is a dependent, the recovery may be substantial, whereas if the decedent was a child or unmarried and childless and the only relatives are parents, the recovery may be small, because most parents receive little economic value from their children. In contrast, if the claimant is the estate, the recovery may include all of the lost economic value.

Where the claimant is a relative, damages from lost wages may be reduced to reflect the decedent's own consumption spending had the person continued living. Such expenses may be relatively small if the claimant is a spouse with children under the theory that much of the decedent's income would have been spent to support the dependents. If decedents have no children or if there is another earner in the family, offsets for the spending of decedents on themselves may be higher.

Shortened Life Expectancy

An important issue is whether a plaintiff may recover compensation for shortened life expectancy caused by an injury. This issue may arise, for example, in medical malpractice cases in which a doctor fails to diagnose and treat a condition or where a surgeon fails to remove a medical device used during surgery.

A related issue is whether dependents in a wrongful-death action may recover economic damages for support the decedent would have provided had the decedent lived—that is, whether such damages can be recovered over the remainder of the decedent's expected lifetime, had he lived. Quantifying such damages requires a projection of the decedent's life expectancy using the methods discussed above.

Acknowledgments

The authors would like to express their gratitude for the substantial contributions of Eduardo Hurtado and Neill Norman in the development of this reference guide.

Glossary of Terms

avoided cost. Cost that the plaintiff did not incur as a result of the harmful act.

Usually it is the cost that a business would have incurred in order to make the higher level of sales it would have enjoyed but for the harmful act.

before-and-after damages. A damages model based on the assumption that the period prior to the challenged conduct (or at some point after the challenged conduct) provides a reliable basis for conditions in the but-for world, absent the challenged conduct.

but-for analysis. Description of the plaintiff's economic situation but for the defendant's harmful act. Damages are generally measured as but-for value less actual value received by the plaintiff.

capitalization. Calculation of today's equivalent to a past dollar to reflect the time value of money and risk. If the capitalization rate is r , the capitalization factor applicable to one year in the past is:

$$\text{capitalization factor} = (1 + r).$$

The capitalization for two years is the capitalization factor squared; for three years, it is the capitalization factor cubed, and so on for longer periods.

compound interest. Interest calculation giving effect to interest earned on past interest, as opposed to simple interest, which does not. Simple interest at interest rate r over a number of periods t and on a principal amount P is computed as $P(1 + rt) - P$, while compound interest earned over the same period is computed as $P(1 + r)^t - P$. Thus, for example, given an interest rate of 20%, the compound interest earned on a lost dollar of income three years earlier is computed as $(1 + 0.20)^3 - 1 = \$0.73$, while simple interest is $(1 + 0.20 \times 3) - 1 = \0.60 .

difference-in-differences. A regression methodology that seeks to identify the impact of the challenged conduct by comparing the change in outcomes (e.g., prices or market share) over time between a group (the treatment group) exposed to the challenged conduct, and a second group (the control group) not exposed to the challenged conduct.

discount rate. Rate used to discount future cash flows.

discounting. Calculation of today's equivalent to a future dollar to reflect the time value of money and risk. If the discount rate is r , the discount factor applicable to one year in the future is:

$$\text{discount factor} = 1/(1 + r).$$

The discount for two years is the discount factor squared; for three years, it is the discount factor cubed, and so on for longer periods.

Reference Guide on Estimation of Economic Damages

earnings. Economic value received by the plaintiff. Earnings could be salary and benefits from a job, cash flows from a business, royalties from licensing intellectual property, or the proceeds from a one-time or recurring sale of property. Earnings are measured net of costs. Thus, lost earnings are lost receipts less costs avoided. Note that the term earnings can also be used to mean the accounting concept of net income, which is not a purely economic concept.

fixed cost. Cost that does not change with a change in the amount of products or services sold.

prejudgment interest. Interest on losses occurring before trial.

present value. Value today of money due in the past (with interest) or in the future (with discounting).

price erosion. Effect of the harmful act on the price charged by the plaintiff. When the harmful act is wrongful competition, as in intellectual property infringement, price erosion is one of the ways that the plaintiff's earnings have been harmed.

regression analysis. Statistical technique for inferring stable relationships among quantities. For example, regression analysis may be used to determine how costs typically vary when sales rise or fall.

variable cost. Component of a business's cost that would have been higher if the business had enjoyed higher sales. See also avoided cost.

Prologue to the *Reference Guide on Exposure Science and Exposure Assessment*, the *Reference Guide on Epidemiology*, and the *Reference Guide on Toxicology*

Courts are sometimes faced with cases that involve toxic substances that cause disease rather than traumatic injury. Those cases often raise difficult questions of factual causation for which scientific evidence is important or essential. The next three reference guides address the disciplines of exposure science and exposure assessment, epidemiology, and toxicology, each of which often plays an important role in such cases.

The discipline of exposure science and exposure assessment focuses on assessing what the dose of a particular substance might have been under a defined set of circumstances. Simply stated, “dose” refers to the amount of a toxic substance that has entered the body. Dose comprises both the magnitude of exposure and the duration of the exposure. Exposure scientists describe the conditions that determine the extent to which a toxic substance enters the body of a person exposed to that substance. For example, if a toxic substance is poorly absorbed through the digestive tract but passes easily into the bloodstream from the lungs, exposure science endeavors to explain the differences between swallowing or inhaling a given quantity of that substance.

The discipline of toxicology focuses on determining the “response” to a dose—that is, both whether an agent causes a specific type of adverse (toxic) effect and how much of a particular substance it takes to cause that effect. Because of the obvious ethical limitations of studying known or suspected toxic substances directly in humans, toxicology relies on a combination of experimental studies using animals (called *in vivo* studies) and human-derived cells and/or tissues (called *in vitro* studies). The experimental setting enables toxicologists to carefully control and define dose and exposure. Much of toxicology seeks to understand the mechanism by which a substance causes adverse biological effects.

The discipline of epidemiology is focused (in this context) on evaluating whether the exposure of a population of individuals to a particular substance is associated with a change (usually an increase) in the occurrence of a specific

disease. Epidemiology faces the same ethical constraints against human experimentation as does toxicology. As a result, except for randomized clinical trials of pharmaceutical agents with potential benefits thought to outweigh their potential risks, epidemiologic studies are observational rather than experimental. Observational epidemiologic studies cannot control exposure and dose as toxicological studies can. But epidemiologists, unlike toxicologists, are able to study groups of human beings rather than human tissues or nonhuman animals.

Each of these disciplines may be brought to bear on a particular scientific question concerning a suspected toxic substance, and they interrelate in important ways. Epidemiologists rely heavily on exposure science to help define the dose and duration of exposure to a substance that individuals in a study population received. Toxicologists also rely on careful assessment of dose. Toxicological information about a substance's mode or mechanism of action helps in assessing whether an observed epidemiologic association reveals a causal relationship between exposure and disease. Conversely, epidemiologic information about a substance's association with disease in human populations helps in assessing whether toxicological effects seen in animal or in vitro studies may validly be extrapolated to people.

Thus, these three reference guides, while independent, share many common elements. As much as possible, the authors have attempted to use common terms and definitions, although the reader should understand that usage may differ across disciplines. Not surprisingly, there will be some redundancy found among the three reference guides. But such redundancy is perhaps useful to highlight critically important elements that are necessary when assessing the scientific strengths and limitations of causal claims in litigation involving toxic substances and disease. The authors of these three reference guides encourage careful consideration and integration of the key scientific elements highlighted in all three guides: the *Reference Guide on Exposure Science and Exposure Assessment*, the *Reference Guide on Epidemiology*, and the *Reference Guide on Toxicology*.

Reference Guide on Exposure Science and Exposure Assessment

M. ELIZABETH MARDER AND JOSEPH V. RODRICKS

M. Elizabeth Marder, Ph.D., is Staff Toxicologist, California Department of Public Health (previously Senior Environmental Scientist, California Environmental Protection Agency's Office of Environmental Health Hazard Assessment), and Assistant Adjunct Professor of Environmental Toxicology, University of California at Davis.

[Note that the views expressed are those of the author and do not represent the California Department of Public Health, the California Environmental Protection Agency, or the State of California.]

Joseph V. Rodricks, Ph.D., is founding principal of ENVIRON (now Ramboll).

[Note that the views expressed are those of the author and do not represent Ramboll.]

CONTENTS

Introduction to Exposure Science, 833

Key Concepts, 834

Practitioners, 835

Relevance to Disease Causation and Risk Assessment, 837

Exposure Science and the Law, 837

Understanding Human Exposure: Key Concepts, 838

Exposure vs. Dose, 838

Characteristics of the Exposed Individual or Population, 839

Exposure Sources, 840

Exposure Routes, 842

Exposure Pathways, 843

Introduction to Exposure Assessment, 847

Exposure Assessment Considerations for Chemical Stressors, 849

Problem Formulation: Why Conduct a Given Exposure Assessment?, 849
Context, 849

Data for Exposure Assessment: Measurements and Models, 850

Measurement Considerations, 853

Model Considerations, 855

Exposure Assessment Methods, 858

Direct Measurement, 858

Indirect Estimation, 860

Exposure Reconstruction, 861

Quantification of Exposure to Chemical Stressors, 863	
Calculation of Exposure, 863	
Integrated Exposure Assessment, 866	
Exposure Assessment of Biologic and Physical Stressors, 868	
How Exposure Science Informs Evaluation of Disease Causation in Individuals, 870	
How Exposure Science Informs Evaluation of Risk in Populations, 873	
Evaluating the Scientific Quality of an Exposure Assessment, 874	
Qualifications of Exposure Scientists or Other Exposure Assessors, 877	
Does the Proposed Expert Have an Advanced Degree or Certification in Exposure Science, Industrial Hygiene, Environmental Health, or a Related Field?, 877	
Is the Proposed Expert Established Professionally in the Field?, 878	
Does the Proposed Expert Have Experience Relevant to the Topic or Method of Interest?, 879	
Appendix A: Units of Exposure and Dose, 881	
Glossary of Terms, 883	
References on Law and Exposure Science, 890	
References on Exposure Science, 890	
Select Guidelines, Reports, and Other Technical Documents, 890	
Select Databases and Tools, 891	
Select Articles, 892	

FIGURES

1. Illustration of source-to-outcome framework for exposure science, 836
2. Examples of exposure sources across a product life cycle, 840
3. Illustration of predictive exposure models and reconstructive pharmacokinetic models in the source-to-outcome pathway, 863

TABLES

1. Weight of Chemical Per Unit Weight of Medium, 881
2. Weight of Chemical Per Unit Volume of Medium, 882

Introduction to Exposure Science

Humans and other organisms regularly come into contact with a wide range of agents that directly or indirectly result in some form of adverse effect or harm. Generally referred to as stressors, these agents are typically chemical, physical, or biologic in nature but may also be psychosocial. Exposure science is a distinct discipline that studies individuals and populations and their behaviors related to contact with stressors, the nature and extent of such contact, and the fate of these stressors in the environment and in organisms over space and time. Though the origins of exposure science date to the ancient Greek physician Hippocrates, modern exposure science is rooted in the industrial-hygiene and radiation health-physics practices of the last century, and the importance of this field is increasingly recognized.¹

Exposure science has a complementary role with epidemiology² and toxicology,³ two disciplines devoted to understanding the inherent potential hazard of a given stressor. This is particularly important in the context of risk assessment, which is a formal process by which the nature and probability of adverse health effects in organisms exposed to a given stressor can be estimated. Given that a dose only results from exposure, exposure science is intrinsically linked with one of the fundamental tenets of human health risk assessment, the toxicological concept attributed to Paracelsus that “the dose makes the poison.” Characterization of exposure plays a key role in environmental epidemiology, as such studies are designed to ascertain whether or not there is an association between exposure to a particular stressor of interest and a health effect in humans. More generally, in order to evaluate whether individuals or populations exposed to a stressor are at risk of harm,⁴ or have actually been harmed, information on toxicity or hazard derived from epidemiological and toxicological studies is needed, as is information

1. In 2012, the National Research Council (NRC) published a seminal report on the field, *Exposure Science in the 21st Century: A Vision and a Strategy*. This report defines exposure science more formally as “the collection and analysis of quantitative and qualitative information needed to understand the nature of contact between receptors and physical, chemical, or biologic stressors.” Nat’l Rsch. Council, *Exposure Science in the 21st Century: A Vision and a Strategy* 20 (2012), <https://doi.org/10.17226/13507>. Following release of this report, the Exposure Science in the 21st Century Federal Working Group (ES21 FWG), consisting of twenty federal organizations, was established to create a common understanding of exposure science, expand partnerships, and drive research across the federal government. In the past decade, federal exposure science programs have rapidly expanded to address increasing demand for exposure data and methods, including programs within the National Exposure Research Laboratory and the National Center for Computational Toxicology of the U.S. Environmental Protection Agency (EPA).

2. See Steve C. Gold et al., *Reference Guide on Epidemiology*, in this manual.

3. See David L. Eaton et al., *Reference Guide on Toxicology*, in this manual.

4. See, e.g., *Rhodes v. E.I. du Pont de Nemours & Co.*, 253 F.R.D. 365 (S.D. W. Va. 2008) (suit for medical monitoring costs because exposure to perfluorooctanoic acid (PFOA) in drinking water allegedly caused an increased risk of developing certain diseases in the future); *In re Welding*

on the exposures incurred by those individuals or populations. This interrelationship is a central theme in a report by the National Academies of Sciences, Engineering, and Medicine (the National Academies) that examines how data being generated today across disciplines can be used in risk assessment applications.⁵

This reference guide emphasizes assessment of exposure to chemical stressors, which serves as a model for considerations of approaches to characterization (e.g., direct measurement, indirect estimation, exposure reconstruction) and quantification of exposure (see sections titled “Exposure Assessment Considerations for Chemical Stressors” and “Quantification of Exposure to Chemical Stressors”). However, some considerations specific to exposure of certain biological and physical stressors are discussed briefly (see “Exposure Assessment of Biologic and Physical Stressors”).

Key Concepts

Exposure science describes the environment, the behavior of stressors in the environment, the characteristics and activities of individuals and populations, and the processes that lead to human contact with and uptake of these stressors. For exposure to occur, the stressor and an individual (or population) need to come together in both space and time. The time of continuous contact between the stressor and individual (or population) is the exposure period. Exposure can be described in terms of intensity or magnitude (how much), frequency (how often) and duration (how long) of contact at an external boundary (those characterized by external exposure surfaces, such as the surface of the skin). For most stressors, both intensity or magnitude and route of exposure are critical characteristics in determining adverse effects. In addition, factors such as the frequency, duration, and timing of exposure (e.g., life-stage considerations, acute versus chronic exposure) are influential in determining adverse effects. These factors depend on the source of the stressor, its transport and fate, its persistence in the environment, and the activities of individuals that lead to contact with the stressor.

Exposure science also extends beyond the exposure event itself (i.e., the point of contact with a stressor) to study and describe the processes that affect the transport and transformation of a stressor, or agent, from its source to a dose at a target internal organ, tissue, or toxicity pathway associated with a disease process. Exposure scientists use a broad range of data and information (empirically derived and modeled) to describe conditions in the real world that could lead to human health risks, typically by conducting an exposure assessment. An exposure

Fume Prods. Liab. Litig., 245 F.R.D. 279 (N.D. Ohio 2007) (exposure to manganese fumes allegedly increased the risk of later developing brain damage).

5. Nat'l Acads. of Scis., Eng'g & Med., Using 21st Century Science to Improve Risk-Related Evaluations (2017) [hereinafter National Academies 2017 Report], <https://doi.org/10.17226/24635>.

assessment is a process of estimating or measuring the magnitude, frequency, and duration of exposure to a stressor, and the characteristics of the population exposed. These exposure assessments, which link exposure to health outcomes and help regulators evaluate various options to manage exposures effectively, can be for current, prospective, and retrospective exposures.⁶ Ideally, an exposure assessment describes the sources, routes, pathways, and uncertainty; describes contact with agents as they occur in the real world at various life stages; and provides data to understand and quantify health outcomes as they occur in various populations.⁷ A source-to-outcome framework, as illustrated in Figure 1, helps to visualize the information and processes important for exposure science, including the major types of experimental and computational approaches to exposure characterization commonly utilized in exposure assessment. Historically, exposure assessment may have ended with the description of the exposure itself (sometimes framed as the potential dose to the body, e.g., prior to crossing a relevant absorption barrier like the skin), but it is now common for practitioners to quantify stressors present in blood, urine, or various tissues of the body as a result of current or past exposure—referred to as biomarkers of exposure—and even to quantify certain biological responses of concern triggered by exposure, referred to as biomarkers of effect.

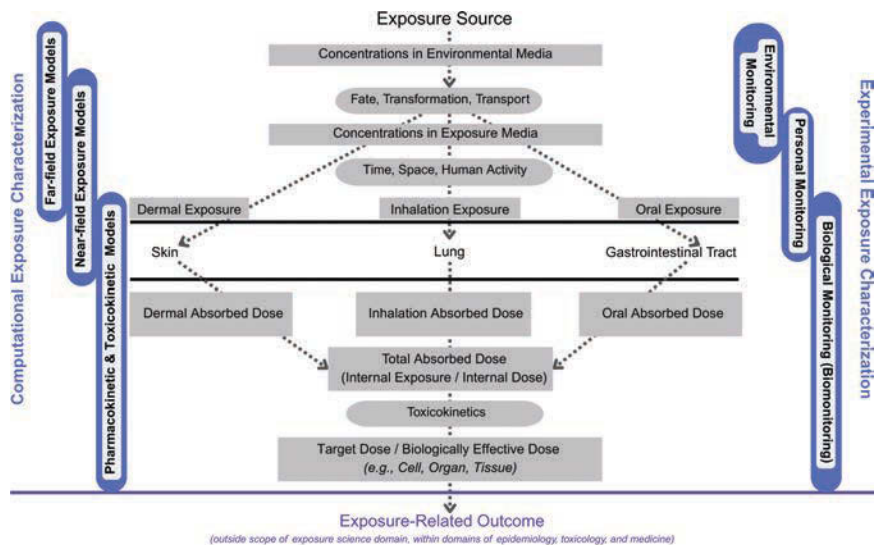
Practitioners

It is important to note that not all exposure assessments are sufficiently complex to require an exposure scientist or similar exposure expert. Indeed, some assessments are relatively simple, with epidemiologists or toxicologists likely capable of estimating exposure in relatively straightforward scenarios, such as consumption of a residue in food or ingestion of a medication. In more complex cases, such as when

6. See, e.g., National Academies 2017 Report, *supra* note 5 (presents recommendations for integrating new scientific approaches into risk-based evaluations; chapter 2 covers advances in exposure science); Nat'l Rsch. Council, Risk Assessment in the Federal Government: Managing the Process (1983), <https://doi.org/10.17226/366> (discusses past efforts to develop and use risk assessment guidelines and evaluates various proposals to modify the risk assessment procedures, including exposure assessment, used by regulatory agencies); Nat'l Rsch. Council, *supra* note 1; U.S. Env't Prot. Agency, Guidelines for Human Exposure Assessment (2019) [hereinafter EPA Guidelines], <https://perma.cc/7XGA-BFKJ> (these present the current policies and practices of exposure assessors within EPA from scoping and problem formulation to presentation of results); Int'l Programme on Chem. Safety, World Health Org., WHO Human Health Risk Assessment Toolkit: Chemical Hazards (2d ed. 2021) [hereinafter WHO HHRA Toolkit], <https://iris.who.int/handle/10665/350206> (section 3.3.4 covers exposure assessment, including a generic roadmap for exposure assessment in the context of human health risk assessment).

7. See, e.g., Nat'l Rsch. Council, *supra* note 1; Linda S. Sheldon & Elaine A. Cohen Hubal, *Exposure as Part of a Systems Approach for Assessing Risk*, 117 Env't Health Persp. 1181, <https://doi.org/10.1289/ehp.080040>; EPA Guidelines, *supra* note 6; WHO HHRA Toolkit *supra* note 6.

Figure 1. Illustration of source-to-outcome framework for exposure science.



Source: Adapted from Figure 2-1 (“Conceptual overview of the scope of and common methods for exposure science”), National Academies of Sciences, Engineering, and Medicine. 2017. *Using 21st Century Science to Improve Risk-Related Evaluations*. Washington, DC: The National Academies Press. <https://doi.org/10.17226/24635>.

exposures result from a stressor moving from sources through one or more environmental media, or when historic exposure profiles must be constructed, it is unlikely that toxicologists or epidemiologists will be able to offer appropriate qualifications, because more advanced modeling and integration of various exposure data are needed to characterize exposures. Given the heterogeneity of exposure science, it should not be surprising that practitioners (exposure scientists as well as related experts) come from a wide range of academic backgrounds. These backgrounds may include environmental health sciences (including some specializing in exposure science), industrial hygiene, environmental and analytical chemistry, chemical and environmental engineering, geology and hydrogeology, toxicology (toxicokinetic applications in particular), epidemiology, and even behavioral sciences (pertaining to those aspects of human behavior that affect exposures). As noted in Figure 1, exposure scientists may address issues of exposure and dose from a source to a biologically effective dose, but these experts are not typically qualified to address health effects of such exposure, which is typically the domain of experts in epidemiology, toxicology, and medicine. Further details on the qualifications of experts are offered in the last section of this reference guide (section titled “Qualifications of Exposure Scientists or Other Exposure Assessors” below).

Relevance to Disease Causation and Risk Assessment

In the evaluation of disease causation for an individual, exposure science is applied to characterize the individual's contact with a stressor, either qualitatively or quantitatively, which is then linked to evidence of disease causation. Exposure assessment is also a major component of risk assessment. In this context, exposure assessments typically must identify and quantify the exposure of populations that are most highly exposed and populations that are most vulnerable, including all relevant exposure pathways (while ideally allowing the pathways to be identified and defined individually), and quantifying inherent uncertainty in the assessment itself.

Exposure Science and the Law

Exposure is the bridge between the presence of a stressor (or agent) in an environment and its ability to cause harm. Therefore, either contact or the plausibility of contact⁸ with a stressor by the individual or population in question must be established before a harmful level of exposure can be definitively determined to have occurred, which is a central challenge in legal cases. Exposure science evidence is particularly relevant in claims of toxic tort or product liability, in which exposure information is critical to effectively assess the liability of defendants in situations where plaintiffs are suing for damages to their health, or the health of their family members, or a group in a class-action claim. Legal actions to protect consumers are usually brought by agencies and groups, including nongovernmental organizations and affected industries, and less frequently by members of the public. Other areas where exposure science is necessary are in formulating a national environmental- or occupational-health-standard control strategy, in developing or updating state or local regulations to protect human health, and in setting safety standards used to protect our food supply or consumer products from contaminants. Not all legal questions concerning human exposures to potentially harmful substances require expert testimony. For example, when the magnitude of exposure is not relevant or is clearly evident (e.g., because a plaintiff

8. See, e.g., *Kitzmiller v. Jefferson*, No. 2:05-CV-22, 2006 WL 2473399 (N.D. W. Va. Aug. 25, 2006) (defendants offered expert testimony that plaintiff's use of liquid cleaning agents containing benzalkonium chloride failed to show that she was exposed to benzalkonium chloride in the air); *Hawkins v. Nicholson*, No. 02-1578, 2006 WL 954654, at *4 (Vet. App. Mar. 2, 2006) (noting that "a veteran who served on active duty in Vietnam between January 9, 1962, and May 7, 1975, is entitled to a rebuttable presumption of exposure to Agent Orange"); *In re Stand 'n Seal*, 623 F. Supp. 2d 1355 (N.D. Ga. 2009) (consumer use of spray-on product allegedly resulted in inhalation exposure to toxic substances, causing respiratory injuries).

was observed to take the prescribed amount of a prescription medicine or was observed being covered in a powder), expert testimony is not indicated. However, expert testimony may be needed in cases involving complex exposure scenarios or in which the magnitude of exposure itself is not a simple question of fact. Such expert testimony—in which a qualitative or quantitative estimate of the magnitude, duration, and frequency of the exposure—is critical to understanding the potential presence or absence of causality between the agent and the health effect, the latter of which typically is based on toxicological, clinical, or epidemiological evidence.

Understanding Human Exposure: Key Concepts

This section of the reference guide is descriptive rather than quantitative. It covers the various physical processes that lead to human exposures to stressors and introduces the terms that exposure scientists apply to those processes, as well as exposure-related characteristics of exposed individuals and populations. Given that exposure science overlaps with other disciplines, many of which use different terms for the same concepts, it is important to note that the definitions used in this reference guide reflect the field of exposure science.

Exposure vs. Dose

Exposure refers to the contact between an agent and the external boundary (exposure surface) of the body for a specific duration. In toxicology, this metric is sometimes also referred to as administered dose or external dose. In this reference guide the general term dose is used to refer to the amount of a stressor (e.g., a toxicant such as benzene) that, over a specified time period, enters the body after crossing an external exposure surface, such as the lungs. In certain contexts, dose may also be referred to as internal exposure, internal dose, or absorbed dose. However, the concept of dose is generally understood to represent the amount of the stressor that gets past the absorption barrier (e.g., lungs) and into circulation (e.g., into blood), a process referred to as uptake,⁹ while the amount of the contaminant that can interact with organs and tissues to cause biological

9. The capacity for a substance to be absorbed via uptake processes is a reflection of its bioavailability. Chemical properties, the physical state of the material to which an individual is exposed, the ability of the individual to physiologically absorb the chemical (e.g., nutritional status, gut flora activity), and even the exposure medium can all affect bioavailability.

effects can be referred to as a biologically effective dose or target dose depending on the context. Note that dose profiles are a function of time, and over time are dependent on factors described for exposure and pharmacokinetics/toxicokinetics (e.g., absorption into the body, distribution throughout the body, metabolism within the body, and elimination from the body).

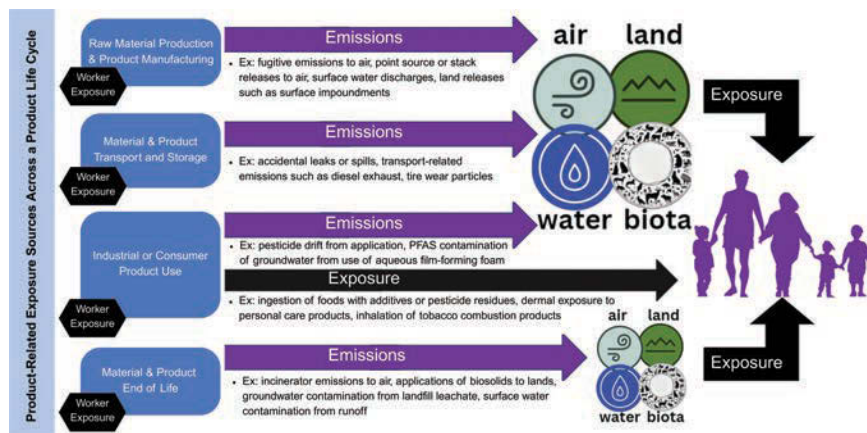
Characteristics of the Exposed Individual or Population

The focus for human exposure science in the context of legal evidence is often the exposed individual, group, or population (sometimes referred to as the receptor or receptor population) and not the sources of the stressor, considering potential contact based on the individual or population's location and behaviors. In this context, understanding the characteristics of human receptors, their behaviors, and the relationship between these factors and exposure is crucial for what is known as a receptor-based approach. There is inherent variability in exposure that occurs because of factors related to human behavior and characteristics that help determine an individual's exposure to a stressor, such as location, occupation, activities within a location, socioeconomic status, consumer preferences, dietary habits, and other lifestyle choices. Such exposure factors include rates of ingestion (e.g., foods, water) or inhalation, factors affecting dermal exposure (e.g., skin surface area), activity factors (e.g., time spent in indoor environment, time spent exercising outdoors), or other factors (e.g., body weight, amount and frequency of consumer product use). Where appropriate, exposure assessors also consider characteristics that might increase exposure or predispose an individual, life stage,¹⁰ specific group, or population to greater health risk. These include factors associated with susceptibility, those in which there is an increased likelihood to be more affected by exposure to a stressor than the general population because of intrinsic factors, such as biological sex, life stage, and genetic polymorphisms. Other factors considered include those associated with vulnerability, such as economic, demographic, social, cultural, psychological, and physical states of the human receptor that influence patterns of exposure to environmental contaminants or alter the relationship between the exposure to environmental contaminants and the health effect of concern (e.g., built environment, access to health care, racism, discrimination). Behaviors relative to life stage can also be particularly influential determinants for exposure, especially for

10. An exposure assessor often needs to establish a dialogue with toxicologists/health scientists to consider whether a specific "window of susceptibility" during a given life stage is important to a particular risk assessment.

infants and toddlers and for a developing embryo/fetus during pregnancy. Note that in most cases, drivers for human activities are complex and cannot be predicted readily. Instead, human activities are often treated as variables described by population distributions based on available (e.g., observational or modeled) data.

Figure 2. Examples of exposure sources across a product life cycle.



Exposure Sources

Human chemical exposure sources can be broadly grouped into near-field sources that are close to the exposed individual, such as consumer product use or occupational exposures, and far-field sources wherein individuals are exposed to chemicals via environmental media following release or use far away. As shown in Figure 2 above, which provides a broad overview of major types of exposure sources across a product's life cycle, there are sources of intended exposure (e.g., dermal exposure to cosmetics or consumption of food additives) and unintended exposure (e.g., accidental releases or spills that reach environmental media such as air or water).

There are many chemicals for which intentional use will lead to human exposures, including substances added to food (and indeed food itself),¹¹

11. The natural constituents of food include not only substances that have nutritional value, but also hundreds of thousands of other natural chemicals. This includes a range of compounds generally considered beneficial for human health (e.g., phytonutrients such as flavonoids and

cosmetics, personal care products, fibers and the colorants added to them, and medical products of many types, including pharmaceuticals; many of these uses are carefully regulated under applicable federal and state laws. In addition to exposure to chemicals that are used to manufacture a product, additional chemical exposures can arise during typical use of a product, such as the anticipated generation of tobacco combustion products from use of cigarettes or of thermal transformation productions from use of electronic nicotine delivery systems. Direct contact with products obviously results in some level of exposure. Here direct contact may mean ingestion via swallowing or other mechanisms, such as inhalation or contact with the skin. There are also some exposures that result from intentional use but are not as well understood, including those related to insertion of some medical devices, such as exposure to silicone and metals from certain types of implants. Generally, however, anticipated exposures resulting from intentional use are more readily quantifiable than those associated with unintended exposures.

Unintended exposures include exposure to known chemicals as well as to unknown chemicals, such as those resulting from unanticipated releases, impurities, or environmental reactions that result in the formation of new substances. Many unintended exposures result from deliberate uses of certain chemicals that, although not intended to lead to human exposures, will inevitably do so. Some amount of a pesticide applied aurally may drift into a nearby residential area, components of food packaging materials may migrate into food, and many types of household products are not intended for direct human ingestion or contact, but exposures nonetheless occur indirectly. As with packaging components, materials used in certain medical devices can contain chemicals that may ultimately leach or volatilize from the material and result in exposure, including plasticizers like phthalates in intravenous tubing and catheters. Ultimately, many exposures to a broad range of environmental contaminants and agents used in occupational settings are unintended; these are often exposures that may not be entirely mitigated by engineering controls and protective equipment, but can also include accidental releases. Unintended exposures are generally more difficult to identify and quantify than are intended exposures; such scenarios require significant expertise to appropriately characterize exposure(s).

polyphenols) as well as those that are harmful, including both toxins (toxic compounds naturally produced by living organisms, such as the cyanogenic glycosides produced by a variety of plants consumed as food, including almonds, cassava, sorghum, and stone fruits), and toxic chemicals taken up from the environment (such as arsenic or lead).

Exposure Routes

An exposure route is the way that a contaminant enters an individual at the point of contact.¹² Inhalation of air containing a substance of interest (vapors as well as particulates) is one of the major routes of exposure.¹³ The internal exposure from inhalation is a function of not only the air concentration of the substance being breathed but also the breathing rate of the individual. The physical form of the substance in air will also influence what happens to the substance during inhalation. For example, chemicals that are in the vapor phase will remain in that physical state and will move to the lungs, where a certain fraction will pass through the lungs and enter systemic circulation. The extent to which different chemical substances pass through the lungs is dependent in large part upon their physical properties, particularly solubilities in both fatlike materials and water. Passage through cell membranes (of the cells lining the lungs) requires that substances have a degree of both fat solubility and water solubility. Certain fibrous materials (including but not limited to asbestos), particulate matter (including but not limited to combustion products), and dusts may also move through the airways and may reach the lungs, but to varying degrees. It is anticipated that at least some such materials will be trapped in the nose and excreted.

Generally, only very fine particles reach the lower lung area and are available for absorption systemically. Some particles may be deposited in the upper regions of the respiratory tract and then carried by certain physical processes to the pharynx and then be coughed up or swallowed. Thus, a variety of inhaled substances can enter the body through the gastrointestinal (GI) tract or the respiratory tract.¹⁴

Ingestion is another major route of exposure to substances in environmental media.¹⁵ Ingestion exposure can occur from intentional consumption of food,

12. EPA Guidelines, *supra* note 6.

13. See, e.g., *Byers v. Lincoln Elec. Co.*, 607 F. Supp. 2d 840 (N.D. Ohio 2009) (welder inhaled toxic manganese fumes); *O'Connor v. Boeing North Am., Inc.*, No. CV 00-0186 DT RCX, 2005 WL 6035256 (C.D. Cal. 2005) (alleged failure to monitor ambient air emissions of radioactive particles); *In re FEMA Trailer Formaldehyde Prod. Liab. Litig.*, No. MDL 07-1873, 2009 WL 2382773 (E.D. La. 2009) (trailer residents exposed to formaldehyde).

14. Joseph V. Rodricks, *From Exposure to Dose*, in *Calculated Risks: The Toxicity and Human Health Risks of Chemicals in Our Environment* 28 (2d ed. 2006), <https://doi.org/10.1017/CBO9780511535451>.

15. See, e.g., *Foster v. Legal Sea Foods, Inc.*, No. CCB-03-2512, 2008 WL 2945561 (D. Md. 2008) (hepatitis A allegedly contracted from eating undercooked mussels); *Winnicki v. Bennigan's*, No. CIV.A. 01-3357 (JAG), 2006 WL 319298 (D.N.J. 2006) (alleged foodborne illness contracted from defendant's restaurant led to renal failure and death); *Palmer v. Asarco Inc.*, No. 03-CV-0498-CVE-PJC, 2007 WL 2254343 (N.D. Okla. 2007) (children allegedly ingested dust and soil contaminated with lead); *Rhodes v. E.I. du Pont de Nemours & Co.*, 253 F.R.D. 365 (S.D. W. Va. 2008) (suit for medical monitoring costs because exposure to perfluorooctanoic acid (PFOA) in drinking water allegedly caused an increased risk of developing certain diseases in the future).

water, other liquids, and certain other substances designed for consumption, such as medicines or supplements. Ingestion exposure can also occur from the intentional or inadvertent nondietary ingestion¹⁶ of soil, dust, or other substances, or of chemical residues on surfaces or objects that are contacted via hand-to-mouth or object-to-mouth activity (especially relevant for young children); this includes incidental ingestion of substances such as cosmetics, personal care products, and certain medications applied to the mouth or lips. They are swallowed, enter the GI tract, and to greater or lesser degrees are absorbed into the bloodstream at various locations along that tract. This is often referred to as the oral route of exposure.

Dermal exposure, the remaining major route of exposure for substances in products and the environment, reflects contact with the largest organ of the body, the skin.¹⁷ The skin is composed of two layers, with a thin outer layer of squamous keratinocytes (called the epidermis or the stratum corneum) that is highly hydrophobic and provides the protective barrier function of skin. Beneath the epidermis is a much thicker living layer of cells including blood vessels, nerves, hair follicles, and sweat glands. As with the GI tract and the lungs, substances are absorbed through the skin to greater or lesser degrees, depending on their physical and chemical characteristics. Other factors also influence dermal absorption, including the specific surface involved, as the skin is not uniform in thickness or in its composition of skin surface fluids, its temperature, and whether occlusion (e.g., trapping material between skin and clothing) has occurred. The uptake of chemicals through these two skin layers is governed by diffusion, and therefore regulated by Fick's law, which states that the rate of diffusion across a barrier will be directly proportional to the concentration gradient. Regardless of exposure route(s) of interest, it is important to note that, depending on the nature of the stressor present at an absorption boundary, in some cases harm can occur directly within the respiratory or GI tracts or on the skin before absorption occurs.¹⁸

Exposure Pathways

An exposure pathway reflects the course a stressor takes from the source to the point at which it reaches the human receptor of interest (e.g., individual, group,

16. Such nondietary ingestion occurs to some extent in all individuals. However, there are individuals for whom this can be more considerable, such as those with pica behavior or similar conditions.

17. See, e.g., *United States v. Chamness*, 435 F.3d 724 (7th Cir. 2006) (evidence that methamphetamine and the ingredients used in its manufacture are toxic to the eyes, mucous membranes, and skin supported sentencing enhancement for danger to human life).

18. Rodricks, *From Exposure to Dose*, *supra* note 14.

population).¹⁹ Exposure pathways analysis allows the identification of all the routes by which stressors from a given source may enter the body, because it identifies all relevant media of human contact into which the stressors migrate. To ensure thoroughness in the assessment, all conceivable pathways should be explicitly identified, with the understanding that ultimately some pathways will be found to contribute negligibly to the overall exposure. The simplest pathways are those described as direct exposure. For example, a substance, such as a non-caloric sweetener or an emulsifier, once added to food follows a simple and direct pathway to the people who ingest the food. The same can be said for pharmaceuticals, cosmetics, and other personal care products. Calculating exposure to such substances, as shown below in the section titled “Quantification of Exposure to Chemical Stressors,” is generally a straightforward process. Even in such cases, however, complexities can arise. For example, in the case of certain personal care products that are applied to the skin, there is a possibility of inhalation exposures to any substance in those products that can readily volatilize (move from a liquid to a gaseous state) at room temperatures. One physical characteristic of chemicals that exposure scientists need to understand is their capacity to volatilize. Not all chemicals are readily volatile, but inhalation routes can be significant for those that are volatile, regardless of their sources.²⁰

Indirect pathways of exposure can range from the relatively simple to the highly complex. Many packaging materials are polymeric chemicals—very large molecules synthesized by causing very small molecules to chemically bind to each other (or to other small molecules) to make very long chemical chains. These polymers (polyethylene, polyvinyl chloride, polycarbonates, and others) tend to be physically very stable and chemically quite inert (meaning they have very low toxicity potential). But it is generally not possible to synthesize polymers without very small amounts of the starting chemicals (those small molecules, usually called monomers) remaining in the polymers. The small molecules can often migrate from the polymer into materials with which the polymer

19. See, e.g., *SPPI-Somersville, Inc. v. TRC Cos.*, No. 07-5824 SI, 2009 WL 2612227, at *16 (N.D. Cal. 2009) (groundwater contamination claim was dismissed because there was no current pathway to exposure); *United States v. W.R. Grace Co.*, 504 F.3d 745 (9th Cir. 2007) (affirming exclusion of report, but not expert testimony based on the report, identifying which pathways of asbestos exposure were most associated with lung abnormalities); *Grace Christian Fellowship v. KJG Invs. Inc.*, 2009 WL 2460990, at *12 (E.D. Wis. 2009) (preliminary injunction was denied because the plaintiff did not establish that a complete pathway currently existed for toxicants to enter the building); Nat’l Exposure Rsch. Lab’y, U.S. Env’t Prot. Agency, Scientific and Ethical Approaches for Observational Exposure Studies, Doc. No. EPA 600/R-08/062 (2008), <https://perma.cc/QM3KTAY6>; U.S. Env’t Prot. Agency, *Exposure Factors Handbook* (2011), <https://perma.cc/D8AR-5FT5>.

20. Inhalation exposures to nonvolatile chemicals can occur if they are caused to move into the air as dusts. See Nat’l Rsch. Council, *Human Exposure Assessment for Airborne Pollutants: Advances and Opportunities* (1991), <https://doi.org/10.17226/1544>.

comes into contact. If those materials are foods or consumer products, people consuming those foods or otherwise using those products will be exposed.

Some amount of the pesticides applied to food crops may remain behind in treated foods and be consumed by people.²¹ This last pathway can become more complicated when treated crops are used as feed for animals that humans consume (meat and poultry and farm-raised fish) or from which humans obtain food (milk and eggs). Exposure scientists who study these subjects thus need to understand what paths pesticides follow when they are ingested by farm animals used as food. The same complex indirect pathways arise for some veterinary drugs used in animals from which humans obtain food.²² In the realm of environmental contamination, pathways can multiply, at which point the problem of exposure assessment can become even more complex, as exposure must be aggregated across relevant pathways. Example sources of environmental contamination include air emissions from manufacturing facilities and from numerous sources associated with the combustion of fuels and other organic materials.²³ Similar emissions that reach water supplies, including ground water used for drinking water or for agricultural applications, can result in human exposures through drinking water and food.²⁴ Contaminants of drinking water that are volatile can enter the air when water is used for bathing, showering, and cooking. In particular, contamination of air in homes and other buildings because of the presence of volatile chemical contaminants in the water beneath those structures, a process referred to as vapor intrusion, is of increasing concern.²⁵ Wastes from industrial processes and many kinds of consumer wastes can similarly result in releases to air and water.²⁶ In some cases, emissions to air can lead to the deposition of contaminants in soils and household dusts; this type of contamination is usually

21. Other pathways for pesticide exposure include spraying homes or fields. See, e.g., *Kerner v. Terminix Int'l, Co.*, No. 2:04-CV-735, 2008 WL 341363 (S.D. Ohio 2008) (pesticides allegedly misapplied inside home); *Brittingham v. Collins*, No. CIV. AMD 06-1952, 2008 WL 678013 (D. Md. Feb. 26, 2008) (crop-dusting plane sprayed plaintiff's decedent); *Haas v. Peake*, 525 F.3d 1168 (Fed. Cir. 2008), *overruled by* *Procopio v. Wilkie*, 913 F.3d 1371 (Fed. Cir. 2019) (veteran claimed exposure to Agent Orange).

22. Patricia Frank & James H. Schafer, *Animal Health Products*, in *Regul. Toxicology* 70 (Shayne C. Gad ed., 2d ed. 2001).

23. See, e.g., *Nat. Res. Def. Council, Inc. v. EPA*, 489 F.3d 1250 (D.C. Cir. 2007) (vacating EPA rule for solid waste incinerators); *Kurth v. ArcelorMittal USA, Inc.*, No. 2:09-CV-108RM, 2009 WL 3346588 (N.D. Ind. 2009) (defendant manufacturers allegedly emitted toxic chemicals, endangering schoolchildren); Am. Indus. Hygiene Ass'n, *Guideline on Occupational Exposure Reconstruction* (Susan Marie Viet et al. eds., 2008).

24. *United States v. Sensient Colors, Inc.*, 580 F. Supp. 2d 369, 373 (D.N.J. 2008) (leaching lead threatened to contaminate ground water used for drinking).

25. *Interstate Tech. & Regul. Council (ITRC), Vapor Intrusion Pathway: A Practical Guide-line* (Jan. 2007), <https://perma.cc/N3ED-9DJT>.

26. *Am. Farm Bureau Fed. v. EPA*, 559 F.3d 512 (D.C. Cir. 2009) (EPA outdoor air pollution standards).

associated with nonvolatile substances. Some such substances may remain in soils for very long periods; others may migrate from their sites of deposition and contaminate ground water, whereas others may degrade relatively quickly. All such issues regarding the movement of chemicals from their sources through the environment to reach human populations come under the heading of chemical fate and transport.²⁷ Transport concerns the processes that cause chemicals to follow certain pathways from their sources through the environment, and fate concerns their ultimate disposition—that is, the medium in which they finally reside and the length of time that they might reside there. Fate and transport scientists have models available to estimate the amount of chemical that will be present in that final environmental medium, often referred to as the exposure medium.²⁸ Some discussion of the nature of these models is offered below in the sections titled “Exposure Assessment Considerations for Chemical Stressors” and “Quantification of Exposure to Chemical Stressors.” One final feature of pathways analysis that should be noted concerns the fact that some chemicals degrade rapidly when they enter the environment, others slowly, and some not at all, or only exceedingly slowly.

The study of environmental persistence of different chemicals is a significant feature of exposure science; its goal is to understand the chemical nature of the degradation products and the duration of time the chemical and its degradation products persist in any given environmental medium. Most inorganic chemicals are highly persistent; metals that become contaminants may change their chemical forms in small ways (lead sulfide may convert to lead oxide), but the metal persists forever (although it may migrate from one medium to another). Most organic chemicals degrade in the environment as a result of their exposure to light, to microorganisms present in soils and sediments, and to other environmental substances. But many organic substances (e.g., polyfluorinated substances such as perfluorooctanoic acid (PFOA), polychlorinated biphenyls (PCBs), and chlorinated dioxins such as dichloro-diphenyl-trichloroethane (DDT)) are quite resistant to degradation and may persist for unexpectedly long periods (although even these ultimately degrade). Exposure assessors also need to be aware of the possibility that the degradation products of certain chemicals may be as toxic, or even more toxic, than the chemicals themselves. The once widely used solvents trichloroethylene and perchloroethylene (tetrachloroethylene) are commonly found in ground water. Under certain conditions, these compounds degrade by processes that lead to the replacement of some chlorine atoms by hydrogen atoms. One product of such degradation is the more hazardous chemical vinyl

27. The common phrase used by exposure scientists is “fate and transport.” In fact, transport takes place and has to be understood before fate is known.

28. In the context of exposure science, the term “final” refers to the medium through which people become exposed. A chemical may in fact continue to move to other media after that human exposure has occurred.

chloride (vinyl chloride monomer or chloroethene), the presence of which in drinking water should not be ignored in an assessment.

A description of pathways is the critical step in characterizing exposure and, especially for environmental contaminants, must be done with thoroughness. Are all conceivable pathways accounted for? Have some pathways been eliminated from consideration, and if so, why? Are any environmental degradation products of concern? Only with adequate description can adequate quantification (see section titled “Quantification of Exposure to Chemical Stressors” below) be accomplished.

Introduction to Exposure Assessment

The primary objective of an exposure assessment is to estimate exposure to the stressor(s) of concern to the human receptor. The completion of an exposure assessment provides the information needed (the amount or concentration and duration of exposure) by epidemiologists and toxicologists, who will have information on the adverse health effects of the chemicals involved and on the relationships between those effects and the resulting dose.²⁹ It is important to note that exposure assessment, and underlying exposure data, can directly contribute to sources of error in epidemiological studies, as discussed in the section titled “Sources of Error in Epidemiologic Studies” of the *Reference Guide on Epidemiology*, in this manual.³⁰ Examples of such sources of error are included in that reference guide and include issues such as methods for imputation of measurements below a method limit of detection (see also “Data for Exposure Assessment: Measurements and Models” below).

As discussed in the next section, exposure assessments can be directed at exposures that occurred in the past, those that are currently occurring, or those that will occur in the future should certain actions be taken (e.g., the entry of a new product into the consumer market or the installation of new air pollution controls). Some aspects of exposure assessment may vary, based on application. Note that various regulatory programs publish their own guidance documents.³¹ As a field, exposure science has moved beyond the classical, source-oriented

29. See Steve C. Gold et al., *Reference Guide on Epidemiology*, and David L. Eaton et al., *Reference Guide on Toxicology*, in this manual. See also, e.g., *White v. Dow Chem. Co.*, 321 Fed. App'x 266 (4th Cir. 2009) (plaintiff must show more than possible exposure; must show concentration and duration); *Anderson v. Dow Chem. Co.*, 255 Fed. App'x 1 (5th Cir. 2007) (lawsuit dismissed because uncontested data showed that magnitude and duration of exposure was insufficient to cause adverse health effects).

30. See Steve C. Gold et al., *Reference Guide on Epidemiology*, in this manual, for a discussion on sources of error in epidemiology studies.

31. See, e.g., U.S. Env't Prot. Agency, *Exposure Factors Handbook* (2011), <https://perma.cc/D8AR-5FT5>.

approaches (those that measure a stressor at its source and then estimate how much reaches an individual or population) to incorporate more human receptor-oriented approaches (those that measure or model a stressor at the individual or population level). Exposure assessments have also become more sophisticated given advances in technologies: These range from dramatic improvements in sensitivity and specificity of targeted analytical methods and incorporation of nontargeted analytical methods to the use of remote and personal sensors/dosimeters in the measurement of exposure, as well as more robust modeling approaches that have allowed for development of much more refined exposure profiles.

Exposure assessment is generally intended to answer the following questions:

- Who has been or could become exposed to a specific stressor(s) arising from one or more specific sources? Is it the entire general population, or is it a specific subpopulation (e.g., those residing near a certain manufacturing or hazardous waste facility; infants and children; or workers)?³²
- What specific stressors comprise the exposures?
- What are the pathways from the source of the stressor to the exposed population? Pathways include direct product use, or those so-called indirect pathways in which the stressor moves through one or more environmental media to reach the media to which people are exposed (air, water, foods, soils, and dusts). Understanding pathways is necessary to understanding exposure routes (below) and quantifying exposures.
- By what routes are people exposed? Routes include ingestion, inhalation, and dermal contact.³³ Identifying exposure routes is important because those routes affect the magnitude of ultimate exposures and because they often affect health outcomes.
- What is the magnitude and duration of exposure incurred by the population of interest? Magnitude (often reported as concentration, amount, or intensity) is the amount of a stressor entering the body or contacting the surface of the body, usually over some specified period of time (often over a day).³⁴ Duration refers to the number of days over which exposure occurs. Has exposure changed over time? Note that

32. See, e.g., *Hackensack Riverkeeper, Inc. v. Del. Ostego*, 450 F. Supp. 2d 467 (D.N.J. 2006) (river and bay users alleged that hazardous waste runoff and emissions polluted the water); *Bradford v. Citgo Petroleum Corp.*, 237 So. 3d 648 (La. Ct. App. 2018) (affirming jury award where the trial court was given “significant circumstantial evidence” that tied plaintiffs living/working/socializing in areas around the CITGO facility to exposures to the subject chemical spills and emissions).

33. Note that additional routes of exposure (e.g., injection, ocular exposure) may be relevant for some pharmaceuticals, diagnostics, and medical devices.

34. Shorter periods of time may be used when the concern is very short-term exposures to chemicals that have extremely high toxicity.

exposures can be intermittent or continuous and can be highly variable, especially for some air contaminants.

Exposure Assessment Considerations for Chemical Stressors

Problem Formulation: Why Conduct a Given Exposure Assessment?

What an exposure assessment is intended to inform will guide the direction of the assessment itself, including the level of the complexity involved (e.g., screening-level assessment). A specific goal of a given exposure assessment might be identifying exposed individuals, groups, or populations, screening chemicals for potential exposure, or identifying source(s) of contamination. Exposure assessments are carried out for a wide variety of reasons, including use in risk assessment, identifying trends in measurements, mitigation efforts, regulatory decision making, priority setting, and epidemiological studies. For example, an assessment completed as part of a regulatory action may involve certain types of legal considerations (e.g., statutory requirements, mandates under a regulator program), whereas an assessment completed as part of an epidemiologic investigation would not.

Context

The development of exposure estimates can follow different methodological approaches depending on the needs of the evaluation to be carried out. For example, if the interest is directed toward personal exposure, the level of exposure should be measured or estimated at the point of contact with the subject (e.g., dermal, inhalation exposure). If exposure is associated with a specific scenario, as is common with near-field exposures, the exposure can be estimated or measured and, to subsequently refine the estimate, combined with information relating to the frequency and duration of the exposure. An exposure estimate can also be carried out with a retrospective approach (e.g., development of historical profile of exposure, reconstruction of exposure from levels of the chemical or metabolites in an individual) or prospective approach (e.g., estimate of present or future exposures). The prospective approach is typically used for regulatory purposes. The retrospective approach can be used, for example, to characterize exposure based on measures or estimates of concentrations in environmental media but can also be used to estimate past occupational exposures for an individual or a population in relation to specific tasks or professional roles held. Such

retrospective exposure assessments are common in toxic tort claims. Note that regardless of whether a prospective or retrospective approach is warranted, a traditional exposure assessment can be designed to consider an exposure resulting from a single identified source and pathway, or from combined exposures to a single chemical resulting from multiple sources, and/or multiple routes and pathways (typically referred to as aggregate exposure).

An aggregate exposure assessment approach is commonly used when humans can be exposed to a single contaminant in various ways. For example, if residues of the same pesticide could be found on multiple foods, in water, and/or in products used in and around the home, then an individual might have exposure via dermal contact, inhalation, ingestion, and other routes. Note that a cumulative exposure assessment can also be undertaken if there is a need to estimate exposure to multiple stressors by multiple routes and/or multiple pathways. Such cumulative exposure assessments are typically conducted for contaminants that produce toxic responses by the same mode of action, or when a population in a specific location is exposed to a variety of stressors.³⁵ The United States Environmental Protection Agency (EPA) has developed guidance for these assessments,³⁶ particularly because cumulative, community-based assessments that can characterize exposures or risks that disproportionately and unfairly affect certain communities are a cornerstone of assessments of environmental justice issues.

Data for Exposure Assessment: Measurements and Models

As discussed previously, exposure science has moved from a focus on classical source-oriented approaches to incorporate more receptor-oriented approaches. Exposure measures have significantly evolved from limited descriptions of a chemical or other stressor in environmental media, food, consumer products, or biological specimens. Biomonitoring data in particular are increasingly available, both in terms of chemical coverage and diverse populations. Biomonitoring refers to the measurement of cellular, biochemical, analytical, or molecular measures obtained from biological media (e.g., tissues, cells, fluids) that can be used to monitor the presence of (1) a chemical in the human body, (2) biological

35. See, e.g., *Diné Citizens Against Ruining Our Env't v. Haaland*, 59 F.4th 1016 (10th Cir. 2023) (finding that Bureau of Land Management acted arbitrarily and capriciously by failing to account for the cumulative impact of hazardous air pollutant emissions from the more than 3,000 wells; these emissions could result in long-term exposure for individuals living in or visiting the San Juan Basin).

36. See, e.g., U.S. Env't Prot. Agency, *Exposure Assessment Tools by Tiers and Types—Aggregate and Cumulative*, <https://perma.cc/LC2L-28K5>.

responses, or (3) adverse health effects.³⁷ These data are typically combined with pharmacokinetic models that simulate the distribution and movement of chemicals within a living system to reconstruct or estimate the amount of chemical to which a person was exposed. Biomonitoring data are particularly useful for reconstruction of aggregate and cumulative exposure as these measures reflect exposure to chemical(s) of interest from all sources, routes, and pathways as well as uptake and accumulation. However, that does mean that these measures are not source or pathway specific and, depending on the scenario and other available data, it may not be possible to identify specific sources or routes of exposure—which limits utility for certain types of assessments.

Biomarkers relevant to exposure assessment include biomarkers of exposure, biomarkers of effect, and biomarkers of susceptibility. The type of biomarker and the biological matrix in which it can be measured are dependent not only on the chemical of interest and its properties, but on inter-individual variability of intrinsic and extrinsic factors that can modify relevant processes, such as metabolism and excretion (e.g., cigarette smoking induces the activity of human enzymes involved in metabolism of various exogenous substances; genetic polymorphisms in such enzymes).

Biological half-life, the length of time required for the concentration of a particular substance to decrease to half of its starting dose in the body, can be highly variable across species and among individuals of the same species. Consideration of biological half-life is particularly important when evaluating biomonitoring data as some chemicals, such as bisphenol A (BPA) and analogues, have an overall biological half-life on the order of hours, while other chemicals, such as dioxins and polychlorinated biphenyls (PCBs), range from years to over a decade. Matrix-specific biological half-life can further complicate this. For example, lead is cleared relatively quickly from the blood and soft tissues with a half-life of one to two months, but is cleared much more slowly from bones, with a half-life of years to decades. As such, biomarkers can be of varying utility depending on the chemical of interest, window of exposure, and sampling period. More information on application of biomonitoring in exposure assessment can be found elsewhere, including the EPA website.³⁸

The EPA has developed working definitions for the types of biomarkers summarized briefly here using examples from pesticide risk assessment.³⁹ Biomarkers of exposure are used to assess the amount of a chemical that is present within the body. Measures include the chemical itself (the most specific biomarker of exposure), chemical metabolites (specificity varies; if a nonspecific metabolite is

37. Nat'l Rsch. Council, *Human Biomonitoring for Environmental Chemicals* (2006), <https://doi.org/10.17226/11700>.

38. U.S. Env't Prot. Agency, *Exposure Assessment Tools by Approaches—Exposure Reconstruction (Biomonitoring and Reverse Dosimetry)*, <https://perma.cc/FCQ4-P7P9>.

39. U.S. Env't Prot. Agency, *Defining Pesticide Biomarkers*, <https://perma.cc/5UT2-XEQ8>.

measured, additional information is needed to determine to which specific chemical the original exposure occurred), and endogenous surrogates. Biomarkers may also include a response within the body that is highly characteristic of a chemical or class of chemicals (this is the least specific biomarker of exposure, as there are many factors that can influence endogenous responses, such as inhibition of butyrylcholinesterase). Biomarkers of effect are indicators of a change in biologic function in response to an exposure (e.g., blood cholinesterase is typically depressed following exposure to organophosphate pesticides), and thus more directly related to insight into the potential for adverse health effects compared with biomarkers of exposure. Biomarkers of susceptibility are factors that result in certain individuals being more sensitive to a given exposure; these biomarkers are therefore more directly related to the potential for adverse health effects than biomarkers of exposure. Examples of biomarkers of susceptibility include genetic factors, nutritional status, lifestyle, and age, among others.

For decades the U.S. Centers for Disease Control and Prevention (CDC) was the leading source of biomonitoring data (with over 300 chemicals measured in biological specimens, such as blood or urine from a nationally representative population sample as part of the National Health and Nutrition Examination Survey). However, there are now thousands of individual studies reporting biomonitoring data available for consideration. Those conducting biomonitoring studies have radically diversified, and they include academic research centers across the United States, state and local biomonitoring programs, and even non-governmental organizations. There are now even direct-to-consumer tests for select environmental contaminants facilitated via lab draws or a home testing kit. Note that not all biomonitoring data will be of sufficient quality for use in exposure assessment; the CDC has developed specific guidance regarding assessing the quality of biomonitoring data. Information is available on the CDC National Biomonitoring Program website.⁴⁰

Advances in technologies—from the use of remote and personal sensor systems to new molecular technologies and computational modeling—allow for rapid development of comprehensive exposure profiles. The use of low-cost sensors is also expanding participatory science, including community-based exposure characterization,⁴¹ which coincides with greater availability of personal data from sources such as social media and wearable devices. Such data are increasingly used for exposure assessment, as they are able to provide insight into exposure on

40. National Biomonitoring Program, U.S. Ctrs. for Disease Control & Prevention, <https://perma.cc/CYC7-EX5F>.

41. See, e.g., Nat'l Inst. Env't Health Sci., Community-Engaged Research and Citizen Science, <https://perma.cc/XGX6-ZV7H>; U.S. Env't Prot. Agency, Participatory Science for Environmental Protection, <https://perma.cc/Y4E3-TSMT>; U.S. Env't Prot. Agency, Citizen Science Opportunities for Monitoring Air Quality (Sept. 2016), <https://perma.cc/JVH3-YP9A>; U.S. Gen. Servs. Admin., Federal Crowdsourcing and Citizen Science Catalog, <https://perma.cc/T4XM-HKS7>.

a timescale of minutes so that peak exposures can be determined. Similarly, remote sensing technology, including satellite sensors, is being applied to map gaseous and particulate air pollution levels on a population and global scale. Nontargeted analytical methods, those that assign chemical formulas and structures to unknown compounds without the use of reference standards or target substance lists, are increasingly used, along with similar “suspect screening” methods, to characterize chemical unknowns in environmental media and in biological media. Such advances in exposure science, particularly those in analytical chemistry, alongside developments of hundreds of new databases and tools,⁴² have also allowed for the untargeted discovery of thousands of chemicals in environmental media, in products, and in humans.

Measurement Considerations

Although at first glance it might seem that direct measurements of environmental concentrations would provide the most reliable data, there are limits to what can be gained through this approach. It is important to recognize that there are a variety of factors influencing accuracy and applicability of available environmental measurement data, including but not limited to proximity to sources, activities of the studied individuals, time of day, season, and weather conditions.

- How can we be sure that the physical samples of media taken are representative of the media of interest for exposure assessment? Standard methods are available to design sampling plans that have specified probabilities of being representative, but they can never provide complete assurance. Generally, when the presence of a chemical is likely to be highly homogeneous, there is a greater chance of achieving a reasonably representative sample than is the case when it is highly heterogeneous. In the latter circumstance, obtaining a representative sample, even when very large numbers of samples are taken, may be unachievable.
- How can we be sure that the samples taken represent the presence of a chemical over long periods? Sampling events may provide a good snapshot of current conditions, but in circumstances in which concentrations could be changing over time, and where the health concerns involve long-term exposures, snapshots could be highly misleading. This type of problem may be especially severe when attempts are being made to reconstruct past exposures, based on snapshots taken in the present.
- How can we be sure that the analytical work was done properly? Most major laboratories that routinely engage in this type of analysis have

42. Over 700 tools are listed in the U.S. Environmental Protection Agency’s ExpoBox (A Toolbox for Exposure Assessors). See <https://perma.cc/T6FB-PS3V>.

developed standard operating procedures and quality control procedures. Laboratory certification programs of many types also exist to document performance. When analytical work is performed in certified, highly experienced laboratories, there is a reasonably high likelihood that the analytical results are reliable. But it is very difficult to confirm reliability when analytical work is done in laboratories or by individuals who cannot provide evidence of certification or of long-standing quality control procedures.

- How are data showing the absence of the chemical to be interpreted? In most circumstances involving possible presence of a chemical in either biological matrices or environmental media, the analysis of some (and sometimes many) of the samples will fail to quantify or detect the chemical of interest. The analytical chemist will often report such samples as nondetect (ND). But an ND should never be considered evidence that the concentration of the chemical of interest (or other target analyte, such as a metabolite) is truly zero. In fact, most chemists will (and should) report that the chemical is below the limit of detection (LOD) or below the limit of quantification (LOQ). Every analytical method has a nonzero limit of detection that refers to the lowest chemical concentration that can be distinguished from a concentration of zero with reasonable confidence. A critical issue is that detection limits vary between assay methods, between laboratories, and even within a laboratory over time.⁴³ Thus, for each sample for which a concentration is reported as ND or below LOD, all that can be known with confidence is that the concentration of the chemical, if present, is somewhere below that limit. If there is clear evidence that the chemical is present in some of the samples (i.e., its concentration exceeds the method's LOD), then it is usually assumed that all the samples of the same medium reported as ND or below LOD will actually contain some level of the chemical. In studies using less sensitive methods to detect a chemical in an environmental or biological matrix, the number of samples with levels below the LOD can be substantial. Studies in which the chemical is detected at relatively low frequencies can lose statistical power by omitting samples below the LOD or by imputing samples below the LOD (e.g., assigning a value of zero, the LOD, or some function of the LOD), which can introduce artificial patterns into the data. These approaches can introduce bias into the analysis to varying degrees, based on the sample size, the proportion of observations below the LOD, and the imputation approach used.⁴⁴

43. See, e.g., Dana B. Barr et al., *A Survey of Laboratory and Statistical Issues Related to Farmworker Exposure Studies*, 114 *Env't Health Persp.* 961 (2006), <https://doi.org/10.1289/ehp.8528>.

44. See, e.g., U.S. Env't Prot. Agency, *Guidance for Data Quality Assessment: Practical Methods for Data Analysis* (2000), <https://perma.cc/SB85-7QKS>. Relevant references from the peer-reviewed

Sampling and measurement are useful, but it is critical to recognize that these are nonetheless limited in important ways. The alternative involves modeling. In fact, a combination of both approaches—one acting as a check on the other—is often the most useful and reliable.

Model Considerations

A model is “a simplification of reality that is constructed to gain insights into select attributes of a particular physical, biological, economic, or social system.”⁴⁵ The simplest model is a conceptual model, from which mathematical processes of the process represented in the conceptual model can be derived (with simplifying assumptions as appropriate). In an exposure assessment, models can be used to extrapolate monitoring data to populations that were not directly analyzed, to reconstruct past exposures, and even to predict future exposures. Models are tools the assessor uses to analyze and characterize processes that are too complex for capturing completely by empirical data or for which empirical data are not available. In the context of chemical exposure assessment, there are a range of models available, including:

- Deterministic modeling (i.e., physical modeling) in which a model describes the relationship between variables mathematically on the basis of knowledge of the physical, chemical, and/or biological mechanisms that are relevant.
 - For example, air particle concentrations in the indoor environment might be modeled by variables such as the air exchange rate, total volume of rooms, and the sources.
- Stochastic modeling (i.e., statistical modeling) in which the statistical relationships are modeled between variables. Such models do not necessarily

literature include the following: Dana B. Barr et al., *A Survey of Laboratory and Statistical Issues Related to Farmworker Exposure Studies*, 114 *Env't Health Persp.* 961 (2006), <https://doi.org/10.1289/ehp.8528>; Haiying Chen et al., *A Distribution-Based Multiple Imputation Method for Handling Bivariate Pesticide Data with Values Below the Limit of Detection*, 119 *Env't Health Persp.* 351 (2011), <https://doi.org/10.1289/ehp.1002124>; Dennis Helsel, *Much Ado About Next to Nothing: Incorporating Nondetects in Science* 54 *Annals Occupational Hygiene* 257 (2010), <https://doi.org/10.1093/annhyg/mep092>; Dennis R. Helsel, *More than Obvious: Better Methods for Interpreting Nondetect Data*, 39 *Env't Sci. & Tech.* 419A (2005), <https://doi.org/10.1021/es053368a>; Dennis R. Helsel, *Less than Obvious: Statistical Treatment of Data Below the Detection Limit*, 24 *Env't Sci. & Tech.* 1766 (1990), <https://doi.org/10.1021/es00082a001>; Jay H. Lubin, *Epidemiologic Evaluation of Measurement Data in the Presence of Detection Limits*, 112 *Env't Health Persp.* 1691 (2004), <https://doi.org/10.1289/ehp.7199>.

45. Nat'l Rsch. Council, *Models in Environmental Regulatory Decision Making* (2007), <https://doi.org/10.17226/11972>.

require fundamental knowledge of the underlying physical, chemical, and/or biological relationships between the variables.

- For example, use of the EPA's Stochastic Human Exposure and Dose Simulation Model (SHEDS) to estimate a product intake fraction for chemicals in consumer products, or use of land use regression modeling for air pollution modeling in which the relation between land use characteristics such as population density, traffic density, green space, and ambient air pollution levels are modeled using statistical regression techniques and measurements.

Perhaps the most widely used models are those that track the fate and transport pathways followed by substances emitted into the air. Knowledge of the amounts emitted per unit of time (usually obtainable by measurement) from a given location (a stack of a certain height, for example) provides the basic model input. Information on wind directions and velocities, the nature of the physical terrain surrounding the source, and other factors needs to be incorporated into the modeling. Some substances will remain in the vapor phase after emission, but chemical degradation (e.g., because of the action of sunlight) could affect media concentrations. Some models provide for estimating the distributions of soil concentrations for those substances (particulates of a certain size) that may fall during dispersion. Much effort has been put into developing and validating air dispersion models.⁴⁶

Similar models are available to track the movement of contaminants in both surface and ground waters. Other types of exposure models currently being employed use geographic information system (GIS) mapping and various regression (e.g., land use regression) and other statistical tools to estimate regional air concentrations for use in exposure models. These approaches, which map locations of high air pollution, have been used in analyses that estimate large-scale population exposures for chemical pollutants such as ozone and fine particulate matter (PM 2.5). The fate and transport modeling issue becomes more complex when attempts are made to follow a chemical's movement from air, water, and soils into the food chain and to estimate concentrations in the edible portions of plants and animals.⁴⁷ Most of the effort in this area involves the use of empirical data (e.g., What does the scientific literature tell us about the quantitative relationships between the concentration of cadmium in soil and its concentration in the edible portions of plants grown in that soil?). This type of empirical information, together with general data on chemical absorption into,

46. *Id.*

47. Exposure scientists specializing in ecological receptors will also use modeling results to evaluate risks to wildlife, plants, and ecosystems.

distribution in, and excretion from living systems, is the usual approach to ascertain concentrations in these food media.⁴⁸

In addition to environmental fate and transport, a variety of human exposure models are readily available for exposure assessment, particularly for assessment of near-field exposures, such as those from consumer products. Availability, and use, of certain types of exposure models has dramatically increased in recent years. This in part results from the advent of the European Union's Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) legislation, which requires manufacturers and importers of hazardous substances eventually included in consumer products to assess exposure and risk from using these products in a chemical safety report, if a substance is manufactured or imported in sufficient quantities. Given the number and range of exposure models available, it is not always readily apparent which models have established validity and which have not. In some cases, however, a given model may have standing with authoritative bodies (e.g., developed or approved by the EPA for a given application). Further complicating matters, it is often the case that a model intended for a specific purpose is being used in a new context that may not have been anticipated by the original developer. An expert should be able to explain why a model is being used and why other possible models are not as suitable.

As models are the basis of many exposure assessments presented as evidence,⁴⁹ if not the major component of such assessments, it is important that an exposure science expert be able to not only describe the model itself and the rationale for its selection but also the scientific basis and underlying assumptions of the model, and the ways in which the model has been validated,⁵⁰ including comparisons to empirical data as available. If empirical data are available, it may also be reasonable to attempt to reconcile the measurement and modeling results and arrive at the most likely values (or range of likely values). Similarly, it may be useful to have the expert describe the likely size of error associated with model results.

48. Nat'l Rsch. Council, *Models in Environmental Regulatory Decision Making* (2007), <https://doi.org/10.17226/11972>.

49. See, e.g., *Milward v. Acuity Specialty Prods. Grp., Inc.*, 969 F. Supp. 2d 101, 108 (D. Mass. 2013), *aff'd sub nom. Milward v. Rust-Oleum Corp.*, 820 F.3d 469 (1st Cir. 2016) (concluding that expert's exposure assessment was admissible where expert used the Advanced REACH Tool (ART) to calculate plaintiff's benzene exposure from using paint that contained mineral spirits); *Brantley v. Int'l Paper Co.*, No. CV 2:09-230-DCR, 2017 WL 2292767 (M.D. Ala. May 24, 2017) (concluding that expert's exposure assessment was admissible where expert reasonably linked the plaintiffs' claims of property damage to the mill's emissions through his AERMOD study).

50. The process of developing and validating a new model or modifying and evaluating an existing model is beyond the scope of this reference guide but U.S. Env't Prot. Agency, *Guidance on the Development, Evaluation, and Application of Environmental Models* (2009), <https://perma.cc/7ARB-67EK>, describes the steps in detail. This specifies a number of steps including credible and objective peer review, corroborating the model by evaluating the degree to which it corresponds to the system being modeled, and performing sensitivity and uncertainty analyses.

Other issues pertaining to the sources and reliability of the data used in the application of a model can be similarly pursued.

Exposure Assessment Methods

The specific methods used in an exposure assessment depend on the exposure assessment questions, availability and feasibility considerations, the intended applications of the exposure assessment (e.g., use in a type of risk assessment, comparison to a reference value), and, in some cases, the regulatory or statutory requirements. For example, an exposure assessment can inform risk screening, enforcement, remediation decisions, or program and policy evaluation. However, even within an exposure assessment, the use of a given method is not mutually exclusive. Accordingly, an assessor may choose to use several methods. Human exposures can be characterized using various approaches: measurements or estimates in the environment (e.g., ambient air concentrations), at the point of human contact (e.g., personal monitoring or sensors), or after contaminants have entered the human body (e.g., biomonitoring). For environmental data, models can be used to estimate human exposure by combining information about environmental concentrations with information about an individual human receptor or a receptor population. Concentrations and receptor information are referred to as exposure factors, which relate to contact with an agent (e.g., inhalation rates, activity patterns, time in microenvironments). Biomonitoring data that reflect an internal dose can also be used to estimate exposure, using “exposure reconstruction,” an approach that typically involves use of pharmac/toxicokinetic models. This section provides an introduction to the three methods most commonly used in exposure assessment: direct measurement, indirect estimation, and exposure reconstruction. The first two methods use information collected or estimated prior to, or at, the point of exposure to predict exposure, while the third, exposure reconstruction, uses information (e.g., measurements of a chemical or metabolite) collected from the body after exposure has already occurred to back-calculate exposure.

Direct Measurement

Direct measurement methods (i.e., point-of-contact methods) use a number of techniques to measure the contact of a person with the chemical concentration in an exposure medium over a specified period of time. Use of direct measurement methods provides an exposure assessor with chemical concentrations or amounts at the interface between the environment and an individual. So long as the techniques are sufficiently accurate, this method is likely to result in the least amount

of uncertainty in estimating exposure concentration, especially over the time period spanned by samples collected. Note that uncertainty can increase in situations in which few measurements are available and short-term sampling data are extrapolated to long-term exposure. Another important consideration is that a direct measurement is not typically source-specific without further data and information. Further, samples from multiple individuals or other units (e.g., households) may be required in order for these measurements to be sufficiently representative of a population of interest. However, these values can be used to validate or verify assessments conducted using an alternate assessment method. Commonly used direct measurement methods include:

- *Passive and active monitors to measure inhalation exposure.* Such monitors are typically compact and located close to the breathing zone of the individual being studied. For passive sampling studies (e.g., use of a diffusion badge to measure formaldehyde) and active sampling studies (e.g., use of a powered pump to draw air through a filter or similar device in order to measure particulate matter), it is often necessary to extrapolate exposure and subsequently dose from the measured concentration as this is dependent on other factors, including breathing rate (see “Understanding Human Exposure: Key Concepts” above).
- *Duplicate diet studies to measure dietary ingestion exposure.* In these studies, individuals will prepare or collect duplicate samples of all the foods they consume during a given period, allowing investigators to measure concentrations of a contaminant in the diet, which are typically normalized to body weight. While this method gives an accurate estimate of ingestion exposure to such contaminants (and can capture information such as changes in concentration owing to preparation processes such as cooking), it is typically cost-prohibitive and difficult for participants. Note that in order to estimate the applied, internal, or biologically effective dose, an exposure assessor would still need to make other assumptions regarding exposure (e.g., leverage toxicokinetic information to predict absorption across the GI tract).
- *Sampling patches, silicone wristbands, whole-body dosimeters, removal methods, and optimal methods to measure dermal contact with chemicals.* Patches or wristbands are materials limited to placement on specific areas of the body to collect a chemical or chemicals of interest. While these sensors are relatively inexpensive and easy to use, they may miss key contact surfaces and introduce error when extrapolating from a small surface area to a larger surface or the whole body. Whole-body dosimeters are coverall suits or full-length cotton garments that are meant to capture any contact with a contaminant in a specific scenario and time frame during which the dosimeter is worn. Removal methods (e.g., rinsing, wiping, tape

stripping) collect contaminants of concern from the skin at a specific interval. Optical methods include fluorescence analyzers, which have been widely used for measurement of various halogenated compounds. These methods all reflect deposition to the skin, so in order to estimate an absorbed or biologically effective dose, an exposure assessor would still need to make additional assumptions (e.g., leverage toxicokinetic information to predict absorption through the skin).

Indirect Estimation

Indirect estimation, sometimes referred to as scenario evaluation, is a method that estimates exposure after developing a set of facts, assumptions, and inferences about a given exposure scenario, using either measurement or estimation of both the amount of a substance contacted by an individual or population and the frequency and duration of this contact. This method typically relies heavily on assumptions about the sources and releases of a chemical of interest, relevant fate and transport mechanisms, and the concentration at the point of contact with the individual or population. This includes identification of the relevant exposure pathway(s) and route(s) of exposure. Similarly, the human receptor of interest (individual or population) must be characterized to generate inputs regarding physiologic processes and activities/behaviors that influence contact with the chemical (especially those that influence the temporal component—i.e., frequency and duration).

Estimates that quantify intake and uptake rates of the chemical for a given individual or population may also be needed. This process often involves development of a conceptual model, a diagram, or a description that lays out the environmental pathways and routes of exposure in the context of the exposure assessment, distinguishes between what is known/determined and what is assumed/based on default values, and identifies sources of uncertainty. The methods applied for quantification of exposure using indirect estimation will vary depending on the level of complexity and the desired result (e.g., a single value or point estimate of exposure versus a distribution of possible exposure values).

Derivation of a single value for exposure is referred to as a deterministic exposure assessment. These values can be either an estimate of central tendency (e.g., mean, median) that is intended to represent the exposure of the average or typical individual in a population or a high-end exposure within the defined receptor group or population. A high-end estimate is typically a representation of individuals at the upper end of the exposure distribution (e.g., at or above 90th percentile of exposure), which is typically calculated using a combination of high-end and central tendency inputs for given exposure parameters (e.g., intake rate).

Note that, while exposure scenarios can be developed to derive a bounding estimate that captures the highest possible exposure, or a theoretical upper bound, for a given exposure pathway, these are more common in screening-level assessments.⁵¹ In contrast, a probabilistic exposure assessment uses distributions of data for various parameters to generate a distribution of potential exposure estimates. Probability distributions (e.g., Monte Carlo simulation) describe the range of values (probability) that might occur in the given population. The output of a probabilistic assessment is a probability distribution of exposures that reflects the combination of the probability distributions of one or more of the model inputs or parameters. These distributions can help characterize variability, uncertainty, or both, depending on the design of the model and its inputs.

Exposure Reconstruction

Exposure reconstruction uses internal body measurements of a chemical or its metabolite(s) rather than external measurements to estimate dose. Such measurements are known as biomarkers, the cellular biochemical, analytical, or molecular measures obtained via biomonitoring from biological specimens (e.g., blood, urine, saliva) that indicate exposure to a chemical.⁵² Biomonitoring is the process of analyzing such human samples to determine biomarker concentrations. This biomonitoring data can then be combined with computational tools such as pharmacokinetic models to reconstruct or estimate the amount of chemical to which a person was exposed.⁵³ Note that there are circumstances in which biomonitoring data alone may be sufficient to demonstrate exposure, even when there is only a semiquantitative or qualitative determination of the presence of a chemical. Such a situation may arise, for example, when the detection or measured concentration of a chemical is established as an inclusion criterion for a class action. This has occurred recently, with various criteria used across cases dealing with perfluorooctanoic acid (PFOA) exposure, ranging from both detection of PFOA in blood to measurement of PFOA in blood and subsequent comparison to either a specified value or to a concentration that is representative of

51. Note that scenarios developed for bounding estimates are sometimes referred to as “worst case” scenarios in which “everything that can plausibly happen to maximize exposure, dose, or risk does in fact happen. This worst case may occur (or even be observed) in a given population, but since it is usually a very unlikely set of circumstances, in most cases, a worst-case estimate will be somewhat higher than occurs in a specific population.” EPA Guidelines, *supra* note 6. These are useful for screening-level evaluations because if the highest possible exposure is evaluated and found to be not of concern, the more plausible lower exposures will also not be of concern.

52. EPA Guidelines, *supra* note 6.

53. Yu-Mei Tan et al., U.S. Env’t Prot. Agency, Biomonitoring—An Exposure Science Tool for Exposure and Risk Assessment (2012), <https://perma.cc/8QWH-A2GQ>.

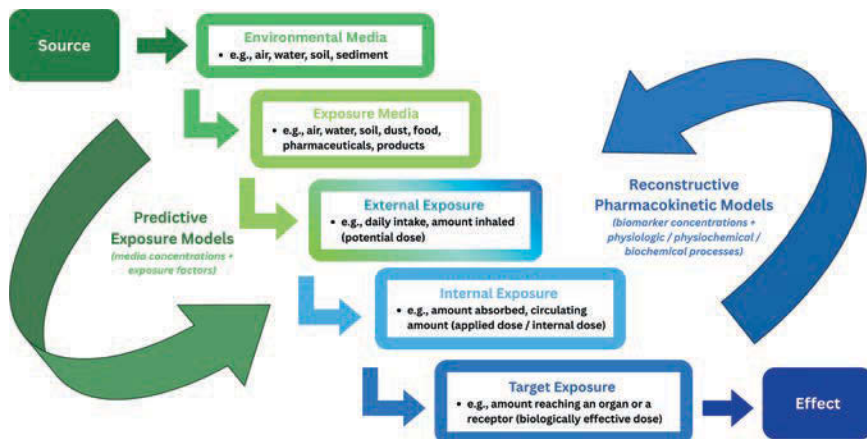
general population exposure.⁵⁴ Pharmacokinetics (PK) or toxicokinetics (TK) characterize the absorption, distribution, metabolism, and excretion of a substance in an organism's body. In exposure assessment, pharmac/toxicokinetic (PK/TK) models use data and mathematical equations to evaluate the fate of a chemical in the body after exposure has occurred. These are highly variable in complexity, with the simplest model being a one-compartment, first-order model that assumes immediate distribution within a single compartment, such as blood.

More complex PK/TK models account for an organism's physiology in their equations and are called physiologically based pharmac/toxicokinetic (PBPK or PBTK) models. These models require parameters to simulate movement and fate of chemicals within the body, considering transfers between tissues and organs, metabolism, and deposition or storage. These parameters can be physiological, physiochemical, or biochemical in nature. Many physiological parameters for these models (e.g., ventilation rates, metabolic rates) are derived from clinical or laboratory studies of animals as well as humans—largely pharmacological studies given the data requirements for such agents. Inter- and intra-species differences in these parameters should be evaluated prior to application.

Figure 3 compares predictive exposure models with reconstructive models. Whereas predictive models use information upstream from the point of exposure, such as environmental media concentrations and exposure factors to estimate an exposure or dose, the reconstructive PK/TK models use information downstream of the point of exposure, such as internal body measurements like biomarker concentrations (e.g., amount absorbed, amount at or bound to a target organ or target biological receptor within the body, and, in some cases, a biomarker of effect) and physiologic processes to estimate the original exposure and internal dose. While reconstructing exposure using biomonitoring data represents definitive exposure to an individual, it is often not possible without further information to identify specific sources or routes of exposure (e.g., inhalation, ingestion, or dermal) that contributed to the level measured in the body.

54. See, e.g., *Hardwick v. 3M Co.*, 589 F. Supp. 3d 832, 869 (S.D. Ohio 2022), *motion to certify appeal granted sub nom. In re E.I. DuPont de Nemours & Co. C-8 Pers. Inj. Litig.*, No. 22-0305, 2022 WL 4149090 (6th Cir. Sept. 9, 2022), *vacated and remanded sub nom. In re E.I. du Pont de Nemours & Co. C-8 Pers. Inj. Litig.*, 87 F.4th 315 (6th Cir. 2023) (Certifies class as “Individuals subject to the laws of Ohio, who have 0.05 parts per trillion (ppt) of perfluorooctanoic acid (PFOA) (C-8) and at least 0.05 ppt of any other PFAS in their blood serum.”); *Benoit v. Saint-Gobain Performance Plastics Corp.*, 959 F.3d 491, 501 (2d Cir 2020) (plaintiffs must show that they have a “physical manifestation of or clinically demonstrable presence of” PFOA in their blood); *Sullivan et al. v. Saint-Gobain Performance Plastics Corp.* 226 F. Supp. 3d 288 (D. Vt. 2016) (requires plaintiffs to show that their blood accumulation levels differ from those in the general population and link blood levels with increased risk of disease and appropriate monitoring).

Figure 3. Illustration of predictive exposure models and reconstructive pharmacokinetic models in the source-to-outcome pathway.



Quantification of Exposure to Chemical Stressors

Calculation of Exposure

It is important to note that quantitative calculations may not always be necessary to sufficiently describe exposure;⁵⁵ however, these are standard practice. The general approaches to characterize exposure are measurements, model estimates,

55. Causation may sometimes be established even if quantification of the exposure is not possible. See, e.g., *Bell v. Abb Grp., Inc.*, No. 13-CV-1338-SMY-SCW, 2015 WL 11439279, at *2 (S.D. Ill. Sept. 1, 2015) (in denial of *Daubert* motion, the court found that the opinion of an expert who conducted a qualitative exposure assessment for an individual, based in part on historical data and exposure modeling reliant on historical analyses of asbestos exposure for a similar occupation, and did not intend to quantify exposure levels, was both supported by evidence and rests on authority and, “as such, cannot be fairly characterized as *mere ipse dixit*”); *Lightfoot v. Ga.-Pac. Wood Prods., LLC*, No. 7:16-CV-244-FL, 2018 WL 4517616, at *20 (E.D.N.C. Sept. 20, 2018) (In denial of *Daubert* motion, the court found that qualitative exposure opinions of experts, both with and without quantitative exposure assessments, are both “relevant and reliable for purposes of supporting their specific causation opinions. See *Westberry v. Gislaved Gummi AB*, 178 F.3d 257], 264 (“[T]his clearly is not a case in which the plaintiff was unable to establish any substantial exposure to the allegedly defective product.”)); *Best v. Lowe’s Home Ctrs., Inc.*, 563 F.3d 171 (6th Cir. 2009) (doctor permitted to testify as to causation based on differential diagnosis).

or a combination of both. Regardless of whether an assessment is informed by measures or models (or both), the calculation of exposure follows the same principles. The simplest calculations relate to situations in which direct exposures occur. For example, consider the case of a substance directly added to food (and approved by the U.S. Food and Drug Administration (FDA) for such addition). Suppose the chemical is of well-established identity and is approved for use in nonalcoholic beverages at a concentration of 10 milligrams of additive for each liter of beverage (10 mg/L).⁵⁶ To understand the amount (weight) of the additive ingested each day, it is necessary to know how much of the beverage people consume each day. Data are available for relevant exposure factors (e.g., rates of food consumption in the general population). Typically, those data reflect average consumption rates and also rates at the high end of consumption.

To make sure that the additive is safe for use, the FDA seeks to ensure the absence of risk for individuals who may consume at the high end, perhaps at the 95th percentile of consumption rates.⁵⁷ Assume that surveys of intake levels for the beverage in our example reveal that the 95th percentile intake is 1.2 L per day for adults. The weight of additive ingested by individuals at the 95th percentile of beverage consumption rate is thus obtained as follows:

$$10 \text{ mg/L} \times 1.2 \text{ L/day} = 12 \text{ mg/day}$$

For a number of reasons, toxicologists often express exposure and dose as weight of chemical per unit of body weight. For adults having a body weight (bw) of, on average, 70 kilograms (kg), the dose of additive is therefore:

$$12 \text{ mg/day} \div 70 \text{ kg bw} = 0.17 \text{ mg/kg bw per day}^{58}$$

Chemical exposure resulting from ingestion of other products containing specified amounts of chemicals are calculated in much the same way. It generally would be assumed that the duration of exposure for a substance added to a food or beverage would be continuous and would cover a large fraction of a lifetime, while for other products, particularly pharmaceuticals, exposure durations will vary widely. It will be useful, before proceeding further, to illustrate calculations for exposures occurring by the inhalation and dermal routes.⁵⁹

56. See *infra* Appendix A for a discussion of units used in exposure science.

57. Vasilios H. Frankos & Joseph V. Rodricks, *Food Additives and Nutrition Supplements*, in Regul. Toxicology 133 (Shayne C. Gad ed., 2d ed. 2001).

58. To gain approval for such an additive, FDA would require that no toxic effects are observable in long-term animal studies at doses of at least 17 mg/kg bw per day (100 times the high-end human intake).

59. See, e.g., *Henricksen v. ConocoPhillips Co.*, 605 F. Supp. 2d 1142, 1164 (E.D. Wash. 2009) (benzene exposure on skin and by inhalation); *Bland v. Verizon Wireless (VAW) L.L.C.*, No. 3:06CV00008-CFB, 2007 WL 5681791, at *9 (S.D. Iowa 2007), *aff'd*, 538 F.3d 893 (8th Cir.

Consider a hypothetical workplace setting in which a solvent is present in the air. Measurement by an industrial hygienist reveals its presence at a weight of 2 mg in each cubic meter (m^3) of air. Data on breathing rates reveal that a typical worker breathes in 10 m^3 of air each 8-hour workday.⁶⁰ Thus, the worker exposure will be:

$$\begin{aligned} 2 \text{ mg/m}^3 \times 10 \text{ m}^3/\text{day} &= 20 \text{ mg/day} \\ 20 \text{ mg/day} \div 70 \text{ kg} &= 0.29 \text{ mg/kg bw per day} \end{aligned}$$

As noted earlier, it is likely that only a fraction of this dose will reach and pass through the lungs and enter the bloodstream. As also noted earlier, if the chemical is a fiber or other particle, its dynamics in the respiratory tract will be different than that of a vapor, with a portion of the inhaled dose entering the GI tract.

In contrast to ingestion and inhalation exposure, dermal exposure often is expressed as the weight of chemical per some unit of skin surface area (e.g., per m^2 of skin). The body surface area of an average (70 kg, ~154 lb) adult is 1.8 m^2 . Thus, consider a body lotion containing a chemical of interest. If the lotion is applied over the entire body, then it is necessary to know the total amount of lotion applied and then the total amount of chemical present in that amount of lotion. That last amount will then be divided by 1.8 to yield the dermal exposure in units of milligrams per square meter. If the chemical causes toxicity directly to the skin, that toxicity dose information also will be expressed in milligrams per square meter. Then risk is evaluated by examining the quantitative relationship between the toxic dose (milligrams per square meter) and the (presumably much lower) human dose expressed in the same units. If the chemical can penetrate the skin and produce toxicity within the body, then a dose determination must include an examination of the amount absorbed into the human body.

One final matter concerning exposure estimation, and subsequently dose estimation, concerns the importance of body size and composition, in particular that of infants and growing children. In matters such as food and water intake, and breathing rates, small children are known to take in these media at higher rates per unit of their body weight than do adults.⁶¹ Thus, when a small child is exposed to a food contaminant, that child will often experience a greater relative exposure to the contaminant than will an adult consuming food with the same

2008) (inhalation exposure to Freon in “canned air” sprayed into water bottle). For a discussion of the importance of assessment of dose as a measure of exposure, see David L. Eaton et al., “Fundamental Principles of Toxicology,” in *Reference Guide on Toxicology*, in this manual.

60. The 24-hour inhalation rate outside the workplace setting is ca. 20 m^3 . The lack of direct proportion to time reflects the fact that breathing rates increase under exertion.

61. See, e.g., Northwest Coalition for Alternatives to Pesticides (NCAP) v. E.P.A., 544 F.3d 1043 (9th Cir. 2008) (dispute over how much lower allowable pesticide levels should be to account for children’s greater susceptibility).

level of contaminant. Children also tend to ingest greater amounts of nonfood items, such as soils and dusts, than do adults. In some cases, nursing mothers excrete chemicals in their breastmilk, which forms a unique source of exposure for nursing infants. The exposure scientist generally conducts separate assessments for children—these assessments take into account the possibility of periods of increased exposure during the developmental period.⁶²

Integrated Exposure Assessment

In many cases, the issue of exposure (and subsequent dose) concerns one chemical in one product and only one route of exposure. But numerous variations on this basic scenario are possible: aggregate exposure involving one chemical in several products or environmental media (e.g., lead exposure from lead-based paint, drinking water in contact with leaded pipes, proximity to airports in which leaded aviation gas is used); cumulative exposures involving either many chemicals in one product or environmental medium (e.g., nitrates and arsenic in drinking water); or many chemicals in many environmental media (e.g., exposure to metals from drinking water, air, consumer products, foods). Even though some exposure situations can be complex and involve multiple chemicals through both direct and indirect pathways, the exposure assessment methods and principles described here can be applied. Exposures occurring by different routes can be added together, or they can be reported separately. The decisions on the final exposure or dose estimates and their form of presentation can be made only after discussions with the users of that information—typically the toxicologists and epidemiologists involved in the risk assessment.⁶³ The exposure or dose metrics emerging from the exposure assessment need to match the exposure or dose metrics that are used to describe toxicity risks.

Further, as discussed in the earlier section on exposure assessment methods, it may be useful to consider high-end or even bounding exposures, not only for exposure to chemicals through foods and consumer products but also in environmental settings. This is because, though a worst case may occur (or even be observed) in a given population, it is usually a very unlikely set of circumstances, and such a worst-case estimate will be somewhat higher than what occurs in a specific population. Therefore, if the highest possible exposure is evaluated and

62. For some substances, susceptibility to toxicity is also enhanced during the same periods.

63. See, e.g., *Am. Farm Bureau Fed'n v. EPA*, 559 F.3d 512 (D.C. Cir. 2009) (challenging EPA's risk assessment for fine particulate matter); *Miami-Dade Cnty. v. U.S. EPA*, 529 F.3d 1049 (11th Cir. 2008) (assessment of risk of wastewater disposal methods to drinking water); *Kennecott Greens Creek Min. Co. v. Mine Safety & Health Admin.*, 476 F.3d 946 (D.C. Cir. 2007) (risk assessment of diesel particulate matter to miners); *Rowe v. E.I. DuPont de Nemours & Co.*, No. CIV. 06-1810 (RMB), 2008 WL 5412912, at *12 (D.N.J. 2008) (risk assessment for proposed class).

found to be not of concern, the more plausible lower exposures will also not be of concern. The EPA's *Exposure Factors Handbook* is one authoritative source of recommended values for general exposure factors needed for estimation of exposure distributions, including such high-end and bounding estimates.⁶⁴

Reviewing the relevant recommended values for direct drinking water ingestion from community sources, a 95th percentile value for drinking water ingestion is approximately 3 L per day across consumers of all ages, with the highest 95th percentile intake of approximately 3.4 L per day for adults aged 21 to 30. Thus, for example, it is possible to assert with relatively high confidence that almost no one in the general population consumes more than 3.5 L of water a day and that almost everyone consumes less, though care should be taken to ensure the appropriate exposure factors are used for the population of interest (e.g., athletes or manual laborers may consume even more water). If an exposure or dose calculation assumes a general population water consumption rate of 3.5 L per day, then the risk estimated for that dose is almost certainly an upper limit on the population risk, and regulatory actions based on that risk will almost certainly be highly protective for the general population. For regulatory and public health decision making, such a precautionary approach has a great deal of precedent, although care must be taken to ensure adherence to scientific data and principles.⁶⁵

This approach becomes problematic, however, if applied to assessments of exposures that may have been incurred in the past by individuals claiming to have been harmed. In such cases, an approach based on attempts to accurately describe the individual's exposure would be necessary. Whatever the case, the exposure scientist must be careful to ensure an accurate description of the exposure concentration (and resulting dose) so that the users of the information can understand whether they have been provided with high-end or bounding estimates, or central tendency estimates (e.g., mean, median) that may describe more typical exposure for a population.

64. U.S. Env't Prot. Agency, *Exposure Factors Handbook* (2011), <https://perma.cc/D8AR-5FT5>. Note: The EPA has developed a companion software tool, ExpoFIRST, that utilizes data from the *Exposure Factors Handbook* in development of user-defined exposure scenarios for a given assessment. ExpoFIRST can then be used to calculate deterministic exposure estimates as point estimates for various receptor populations and life stages. This tool is accessible via the EPA Expo-Box (see <https://perma.cc/T6FB-PS3V>).

65. Nat'l Rsch. Council, *Evolution and Use of Risk Assessment in the Environmental Protection Agency: Current Practice and Future Prospects*, in *Science and Decisions: Advancing Risk Assessment* 26 (2009), <https://doi.org/10.17226/12209>. Those who must comply with regulations that were developed based on a high degree of caution often protest that more accurate assessments should be used as their basis. For several reasons, truly accurate prediction of risk is difficult to achieve (see David L. Eaton et al., *Reference Guide on Toxicology*, in this manual), while predicting an upper bound on the risk is not. At the same time, unless carefully done and described, upper-bound estimates may be so remote from reality that decisions based on them should be avoided.

Exposure Assessment of Biologic and Physical Stressors

In addition to chemical stressors, humans are exposed to a wide range of non-chemical stressors, including biologic stressors (e.g., bacteria, viruses) and physical stressors (e.g., radiation, heat, noise). Much like chemical stressors, these exposures may arise from typical activities, such as consumption of food, or from specific incidents, such as an industrial accident or a natural disaster. Exposure assessment considerations specific to biological and physical stressors are not enumerated here. For many nonchemical stressors, the general conceptual approach, including calculations, to exposure assessment is comparable to those outlined for chemical stressors in the earlier sections, “Exposure Assessment Considerations for Chemical Stressors” and “Quantification of Exposure to Chemical Stressors.” For example, much like a chemical stressor, exposure assessment for radiation can involve various measurements, including the amount of radioactivity, ambient radiation levels in environmental media such as water, soil, or air, and an internal or absorbed dose (energy deposited per unit mass) that a person has received from a radioactive source. However, it is important to note that there can be substantial differences in the type and scope of information and data required, as well as in the exposure measurements and models used to characterize exposure and inform risk assessment for certain nonchemical stressors.

There are considerable differences even in the approach to description of radiation, with four different but interrelated types of units for measuring radioactivity, exposure, absorbed (internal) dose, and effective dose (also referred to as dose equivalent). Typically, the amount of radioactivity refers to the amount of ionizing radiation released by a material, whether it emits alpha particles, beta particles, gamma rays, x-rays, or neutrons. The quantity of radioactive material is expressed in terms of its radioactivity, which represents how many atoms in the material decay in a given time period, with units of measure most commonly reported in becquerels (Bq, international unit, equivalent to one radioactive decay per second) or Curie (Ci, U.S. unit, approximate number of radioactive decays in one gram of radium per second). Such radioactivity can be measured by a variety of detectors, including the commonly used Geiger counter. For measurement of ambient radiation in the environment, other instrumentation, such as pressurized ionization chambers, are best suited, and levels are reported in units such as roentgen per hour (R/h) or coulomb per kilogram (C/kg).

In contrast, an absorbed radiation dose reflects the amount of energy that a radioactive source deposits in an organism through which the radiation passes. Absorbed radiation can be measured using additional specialized instruments, including alarming dosimeters that are typically used to monitor exposure rates and accumulated dose in real time. The conventional unit for absorbed radiation dose is now the international unit Gray (Gy), though the older term rad

(for radiation absorbed dose) may still be used (one Gy is equivalent to 100 rad).⁶⁶ Lastly, the dose equivalent, or effective dose, of radiation is a term that addresses an individual's biological risk given the type of radiation exposure by adjusting the absorbed dose by radiation type and relative organ sensitivity. The conventional unit for dose equivalent is now the sievert (Sv), though the roentgen equivalent man (rem) may still be used in some cases. The dose equivalent is commonly used to set protective levels for groups of people. For example, the annual radiation dose limit for workers in the United States is 0.05 Sv or 5 rem.⁶⁷

Other physical stressors, such as noise and temperature, have their own terminology and approaches, as do the various types of pathogens. As such, appropriate stressor-specific guidance should be considered when evaluating such an assessment. A variety of guidance documents and tools have been developed to support measurement of specific biological and physical stressors, with various federal agencies and nongovernmental organizations having also developed specific guidance for risk assessment of such stressors, particularly for radiation and microbial exposures. For example, to support the assessment of radiation, the EPA and the Oak Ridge National Laboratory (ORNL) have developed a comprehensive software system specifically for the calculation of tissue dose and subsequent health risk from radionuclides in environmental media. Interpretation of these model results requires specialized knowledge of exposure assessment and the integration period of radiological decay. Health physicists, for example, are individuals who have received such specialized training in nuclear physics, radiation biology, radiation detection, radiation chemistry, and other related sciences. Microbial risk assessment has similar specificity, with the U.S. Department of Agriculture/Food Safety and Inspection Service (USDA/FSIS) and EPA jointly developing a guideline for microbial risk assessment with an emphasis on pathogenic organisms in food and water.⁶⁸ Accordingly, exposure assessment for a nonchemical stressor should be conducted by an expert sufficiently familiar with the stressor and appropriate means of exposure and risk assessment.⁶⁹

66. U.S. Ctrs. for Disease Control & Prevention, Nat'l Ctr. for Env't Health, *Measuring Radiation*, <https://perma.cc/P2BF-QZKD>.

67. 10 C.F.R. § 20.1201 (1991), <https://perma.cc/SYP4-Y4LT>.

68. U.S. Env't Prot. Agency (EPA) & U.S. Dep't of Agric./Food Safety & Inspection Serv. (USDA/FSIS), *Microbial Risk Assessment Guideline: Pathogenic Organisms with Focus on Food and Water* (2012), <https://perma.cc/U4QB-RT8J>.

69. *See, e.g., United States v. Mass. Water Auth.*, 97 F. Supp. 2d 155, 185 (D. Mass. 2000), *aff'd*, 256 F.3d 36 (1st Cir. 2001) (The most authoritative risk assessment of one water treatment option was attributed to a scientist with expertise in assessment of microbial risks of drinking water; the court discounted the testimony of another scientist who testified only in general terms about water filtration as the court found him to have "no expertise in water quality issues. He has testified 'maybe a hundred times as a plaintiff's expert in asbestosis cases.'").

How Exposure Science Informs Evaluation of Disease Causation in Individuals

In the evaluation of disease causation for an individual, exposure science is applied to characterize the individual's contact with a stressor, either qualitatively⁷⁰ or quantitatively. The resulting exposure assessment is then linked to evidence of disease causation (e.g., a causal link established using epidemiological and toxicological investigations,⁷¹ but also possibly an association between exposure and outcome identified by the same means). Note that an individual is generally not faced with the burden of demonstrating causality with absolute proof; rather, the legal standard is typically expressed as a need to show that causality is demonstrable by a preponderance of the evidence (e.g., more likely true than not true that the harm observed was caused by the exposure).⁷² Establishing general causality, and providing an expert opinion as to whether it is reasonable to conclude that the exposure caused harm to an individual, is the domain of epidemiology, toxicology, and/or medicine.⁷³ Exposure experts are needed to assess the exposures incurred by an individual and to evaluate the quality and validity of such assessments, while appropriate causation experts are called upon to offer testimony on whether those exposures are of a magnitude sufficient to cause the claimed harm.

70. As discussed in the section titled “Quantification of Exposure to Chemical Stressors” above, causation may sometimes be established even if quantification of the exposure is not possible. See, e.g., cases cited *supra* note 55 (*Bell, Lightfoot, Westberry, & Best*); *Allen v. Martin Surfacing*, 263 F.R.D. 47 (D. Mass. 2009).

71. See Steve C. Gold et al., *Reference Guide on Epidemiology*, and David L. Eaton et al., *Reference Guide on Toxicology*, in this manual, for discussions on epidemiological and toxicological approaches to disease causation, respectively. Regulations and public health actions are usually driven by findings of excessive risk of harm (although sometimes evidence of actual harm).

72. Did the plaintiff incur exposures to the suspect chemical of sufficient magnitude and duration to make it more likely true than not that the chemical, and not some other factor, was the cause of the plaintiff's medical condition? See, e.g., *Barrett v. Rhodia, Inc.*, 606 F.3d 975, 984 (8th Cir. 2010) (No appellant expert was qualified to present evidence that Barrett was exposed to toxic levels of hydrogen sulfide gas as opposed to another harmful exposure; the defense presented such an alternative cause supported by expert testimony. Court noted that “[e]xpert testimony ‘based on possibility or speculation is insufficient [to establish causation]; it must be stated as being at least ‘probable,’ in other words, more likely than not.’ *Fackler v. Genetzky*, 263 Neb. 68, 638 N.W.2d 521, 527–28 (2002).”).

73. Note that a statistical association between exposure and disease does not prove causation, so plausible alternative hypotheses must be eliminated by careful statistical adjustment and/or consideration of all relevant scientific knowledge. Epidemiologic studies can show an association after such adjustment. Such studies, reasonably free of bias and further confounding, provide evidence for but not definitive proof of causation; causal associations can be further strengthened by mechanistic knowledge about how particular agents might produce adverse health effects.

Exposures of concern for individual disease causation can be highly variable in scope, including hazards present in a workplace, contamination of environmental media in a particular location (e.g., release of toxic air contaminants from a facility), and even use of a given product or products. In many cases, exposure to the same stressor may occur simultaneously through multiple media (e.g., air, food, water). This complicates both exposure assessment itself and attribution of exposure to a specific source, which is often relevant. In any such instance, however, assessment of exposure can involve either exposure measurements or modeling (see sections titled “Exposure Assessment Considerations for Chemical Stressors” and “Exposure Assessment of Biologic and Physical Stressors” above). Unless a personal monitoring study has been conducted, at least some modeling is expected regardless of what measured data may be available. Evidence presented should, at minimum, include exposure and dose estimates, and should identify and quantify important source(s) and significant pathway(s) and route(s) of exposure from the source to the individual. As an individual’s exposure is expected to vary over time, often on several different time scales and usually depending on the individual’s pattern of activities and locations, these variables must be described in some level of detail and incorporated into an exposure assessment, in addition to information about the concentration of the exposure in relevant media. This is of critical importance as it is generally expected that there is some degree of association between the intensity and temporal characteristic of the exposure and the biology of the disease; this information can therefore provide information to mitigate a source or end exposure to an individual at risk.

There are two general scenarios in which an individual’s exposure must be characterized. In the first, exposure to a stressor is currently occurring or anticipated to occur. This is typically more straightforward than in the second scenario, most common in toxic tort claims, in which an individual claims to have been harmed by prior exposure to a stressor, often by alleging that some existing medical condition has been caused by exposure, or exposures, occurring in the past. In both scenarios, exposure assessments are used to ascertain whether an individual would be exposed to a stressor, and to what extent.

Assessment of past exposures is especially difficult when considering diseases with very long latency periods.⁷⁴ By the time disease occurs, documentary proof of exposure and magnitude may have disappeared; however, courts regularly deal with evidence reconstructing the past, and assessment of exposure is another

74. In certain tort and insurance litigation, diseases with a long latency period, especially those attributable to continuous or repeated exposure to a substance, may be referred to as “long-tail harm.” See, e.g., *Danaher Corp. v. Travelers Indem. Co.*, 414 F. Supp. 3d 436 (S.D.N.Y. 2019) (resolution of disputes concerning insurance coverage for asbestos- and silica-related claims against a former subsidiary).

application of this common practice.⁷⁵ Depending on the stressor, there may be measurable evidence of past exposure at the time of assessment, such as bio-monitoring data (more likely for a persistent chemical⁷⁶ than for a short-lived chemical). However, such a scenario typically requires reconstruction of exposure, an undertaking that also requires appropriate consideration of factors that may have influenced the disease process, or both the exposure and disease process.⁷⁷ Although there may in some cases be historical measured data available, such as monitoring data in a relevant environmental medium or location (e.g., historic levels of water or air contaminants in a community, which are available for relatively few stressors, mostly chemical or biological), reconstructions of exposure are typically heavily reliant on modeling. Reconstruction of occupational exposures has been a relatively successful pursuit, because often historical industrial hygiene data are available involving the measurement of workplace air levels of chemicals. If it is possible, through the examination of employment records, to understand an individual's job history, it may be possible to ascertain that individual's exposure history and develop a retrospective exposure assessment.⁷⁸ Guidelines for such retrospective occupational exposure

75. Courts have accepted indirect evidence of exposure. For example, differential diagnosis may support an expert's opinion that the exposure caused the harm. *Best v. Lowe's Home Ctrs., Inc.*, 563 F.3d 171 (6th Cir. 2009). On occasion, qualitative evidence of exposure is admitted as evidence that the magnitude was great enough to cause harm. *See, e.g., Westberry v. Gislaved Gummi AB*, 178 F.3d 257 (4th Cir. 1999) (no quantitative measurement required where evidence showed plaintiff was covered in talc and left footprints); *Allen v. Martin Surfacing*, 263 F.R.D. 47 (D. Mass. 2009) (accounts of symptoms at the time of exposure formed the basis for expert's opinion that exposure was high enough to cause harm). And courts have accepted the government's reconstruction of exposure to radiation. *Hayward v. U.S. Dep't of Labor*, 536 F.3d 376 (5th Cir. 2008); *Hannis v. Shinseki*, No. 09-0593, 2009 WL 3157546 (Vet. App. 2009) (no direct measure of veteran's exposure to radiation was possible but VA's dose estimate was not clearly erroneous).

76. *See, e.g., Benoit v. Saint-Gobain Performance Plastics Corp.*, 959 F.3d 491, 501 (2d Cir. 2020) (plaintiffs must show that they have a "physical manifestation of or clinically demonstrable presence of" PFOA in their blood); *Sullivan et al. v. Saint-Gobain Performance Plastics Corp.*, 226 F. Supp. 3d 288 (D. Vt. 2016) (requires plaintiffs to show that their blood accumulation levels differ from those in the general population and link blood levels with increased risk of disease and appropriate monitoring).

77. Confounding factors must be carefully addressed. *See, e.g., Allgood v. Gen. Motors Corp.*, No. 102CV1077DFHTAB, 2006 WL 2669337, at *11 (S.D. Ind. 2006) (selection bias rendered expert testimony inadmissible); *Am. Farm Bureau Fed'n v. E.P.A.*, 559 F.3d 512 (2009) (in setting particulate matter standards addressing visibility, the data relied on should avoid the confounding effects of humidity); *Avila v. Willits Env't Remediation Tr.*, No. C 99-3941 SI, 2009 WL 1813125 (N.D. Cal. 2009) (failure to rule out confounding factors of other sources of exposure or other causes of disease rendered expert's opinion inadmissible); *Adams v. Cooper Indus., Inc.*, No. CIVA 03-476 JBC, 2007 WL 2219212 (E.D. Ky. 2007) (differential diagnosis includes ruling out confounding causes of plaintiffs' disease).

78. *See, e.g., T.W. Armstrong, Exposure Reconstruction, in Mathematical Models for Estimating Occupational Exposures to Chemicals* (Charles B. Keil et al. eds., 2d ed. 2009); Francesca Borghi et al., *Retrospective Exposure Assessment Methods Used in Occupational Human Health Risk Assessment*:

assessments have been published by various agencies, as well as nongovernmental organizations such as the American Industrial Hygiene Association.⁷⁹

A retrospective exposure assessment outside of occupational settings is typically more difficult because there are usually many changes over time—for example, changes in technologies of the pollution sources, and also changes in people’s behavior and location over time.

How Exposure Science Informs Evaluation of Risk in Populations

Exposure assessment is one of the core components of regulatory quantitative risk assessment; the quality of exposure information and the exposure assessment itself are critical to the quality and utility of risk assessment. Risk assessments are typically directed at an existing exposure situation, such as the risks incurred by populations residing in the vicinity of a manufacturing or hazardous waste facility, or at the exposure situation expected if certain regulatory actions are taken. In considering exposures in a population, as opposed to an individual, it is critical to recognize that for any one exposure scenario there is inherently a range of exposures to a given stressor that are experienced by the individuals who compose the population. For example, assuming a toxicant is released from a manufacturing facility into the air in a community, some individuals may have a high degree of contact for an extended period (e.g., factory workers exposed to a substance on the job) while other individuals may have a lower degree of contact for a shorter period (e.g., individuals using a recreational site downwind of the facility). These exposures will be a function not only of the substance’s concentration in the air at the respective points of contact, but also of the characteristics of the individual or population (e.g., inhalation rate is variable across life stage and differs substantially with various activities).

To be useful for population risk assessment, an exposure assessment needs to be capable of identifying and quantifying the exposure of the populations that are most highly exposed and the populations that are most vulnerable. The assessment should include all relevant exposure pathways and allow the pathways to be identified and defined individually (for example, to allow for quantification of exposure from water or from food, as well as total exposure). Exposure assessment also needs to consider background exposures to other stressors that could influence exposure or risk within a population.

A Systematic Review, 17 Int’l J. Env’t Rsch. & Pub. Health 6190 (2020), <https://doi.org/10.3390/ijerph17176190>.

79. Am. Indus. Hygiene Ass’n, Guideline on Occupational Exposure Reconstruction (Susan Marie Viet et al. eds., 2008).

The assessment must also consider inherent uncertainty in the process to determine the level of confidence in the overall risk assessment. Note that in some cases, exposure science can also inform risk to a population in a given scenario through comparison of estimated or measured concentrations of stressors in environmental media (e.g., air, water) to established reference values or standards delineating a level at which no harm is anticipated from exposure. The most common such values are regulatory tolerances for pesticide residues in food, maximum contaminant levels (MCLs) for drinking water contaminants, National Ambient Air Quality Standards (NAAQS), and, for workplace exposure, permissible exposure limits (PELs) or threshold limit values (TLVs).⁸⁰ Such comparisons are intended to determine whether products and environmental media contain substances at concentrations that meet existing regulatory requirements.

Evaluating the Scientific Quality of an Exposure Assessment

Though there is no standard set of criteria for evaluating the quality of an exposure assessment across all possible stressors, some concepts are universally relevant. Regardless of the stressor, all exposure assessments must reflect at least the two principal dimensions of exposure: its magnitude (often reported as concentration, amount, or intensity) and time. Temporal considerations include duration as well as pattern or frequency (e.g., continuous or intermittent exposure), and any relevant windows of susceptibility. Exposure scientists may offer expert testimony regarding exposures to stressors incurred by individuals or populations. In some cases, expert testimony will include description and quantification of concentrations in environmental media or a description and quantification of body burden, while in others, testimony may rely entirely on models rather than measurements of any kind. In each case, such assessments typically include a description of how and when exposures did or could occur, the identities of the chemicals involved, the routes of exposure, the magnitudes of exposure incurred, and the durations of exposure. These may be qualitative, quantitative, or semiquantitative in nature; an

80. PELs are official standards promulgated by the Occupational Safety and Health Administration. TLVs are guidance values offered by an organization called the American Conference of Governmental Industrial Hygienists. See, e.g., *In re Howard*, 570 F.3d 752, 754 (6th Cir. 2009) (challenging PELs for coal mine dust); *Jowers v. BOC Grp., Inc.*, 608 F. Supp. 2d 724, 735–36 (S.D. Miss. 2009) (PELs and TLVs for welders' manganese fume exposure); *Int'l Brominated Solvents Ass'n v. Am. Conf. of Governmental Indus. Hygienists, Inc.*, 625 F. Supp. 2d 1310 (M.D. Ga. 2008) (challenging TLVs for several chemicals); *Miami-Dade Cnty. v. U.S. E.P.A.*, 529 F.3d 1049 (11th Cir. 2008) (MCLs for public drinking water).

exposure assessment does not have to be quantitative to be of use,⁸¹ but typically an estimation of at least semiquantitative⁸² magnitude and duration of incurred exposure is necessary to evaluate plausibility of a causation claim.⁸³

For the purposes of this reference guide, it is assumed that questions regarding disease risk and causation are beyond the bounds of exposure science and are instead within the domains of other sciences, such as epidemiology and toxicology.⁸⁴ However, if the exposure scientist is also an epidemiologist or toxicologist,⁸⁵ or has a relevant medical background, he or she may offer additional testimony on the health risks associated with those exposures or even regarding the question of whether such exposures may have caused disease. Below is a set of questions that exposure scientists should be able to answer, with appropriate documentation and scientific reasoning, to support any given exposure assessment.

- Is the purpose of the assessment clear? Is the exposed individual or population specified?
- Has the source(s) of exposure been identified?
- When did the exposure occur: Past? Present? Future? If exposure is occurring now, will it continue to occur? Under what conditions is future exposure anticipated?
- What is known or assumed about the nature and extent of media contact by members of the exposed population? How has this been ascertained?
- What is the assumed frequency and duration of exposure, and what is the basis for these assumptions?

81. As discussed in the section titled “Quantification of Exposure to Chemical Stressors” above, causation may sometimes be established even if quantification of the exposure is not possible. See cases cited *supra* note 55; *Allen v. Martin Surfacing*, 263 F.R.D. 47 (D. Mass. 2009) (accounts of symptoms at the time of exposure formed the basis for expert’s opinion that exposure was high enough to cause harm).

82. See, e.g., *Rhyne v. United States Steel Corp.*, 474 F. Supp. 3d 733, 761 (W.D.N.C. 2020) (Expert exposure assessment found to be based on sufficient data and to be the product of a reliable methodology, even though estimates are not based on exact data on how often the plaintiff used the product in question at home. The court notes that “[t]o require expert testimony to be based on exact information as to how frequently plaintiff used a product fifty years ago, as Savogran suggests, would effectively prohibit a plaintiff from ever recovering in a latent disease case.”).

83. See Steve C. Gold et al., *Reference Guide on Epidemiology*, in this manual; see also section titled “How Exposure Science Informs Evaluation of Risk in Population” above.

84. See Steve C. Gold et al., *Reference Guide on Epidemiology*, and David L. Eaton et al., *Reference Guide on Toxicology*, in this manual, for discussions on epidemiological and toxicological approaches to disease causation, respectively.

85. See “Qualifications of Exposure Scientists or Other Exposure Assessors” below, which deals with the question of the qualifications of exposure scientists. In many cases, the work of exposure experts is turned over to health experts to incorporate into their evaluation of risk and disease causation. In some cases, usually the less complex ones, exposure assessments may be undertaken by the health experts.

- What are the pathways from the source to the exposed individuals or population(s)? How has it been established that those pathways exist? (Past? Present? Future?)
- What is the concentration of the chemical in the media with which the exposed individual or population comes into contact? (Past? Present? Future?) What is the basis for this concentration: Direct measurement? Indirect estimation? Exposure reconstruction?
- If the concentration is based on direct measurement, what procedures were followed in obtaining that measurement? Was sampling of environmental media sufficient to ensure that it was representative? If not, why is representativeness not important? If a biomarker was measured, how does it relate to the original exposure—is this the stressor itself or a metabolite? Were validated analytical methods used by an accredited laboratory? If not, how can one ensure that the analytical results are reliable?
- If models were used, what is their reliability? (See “Exposure Assessment of Biologic and Physical Stressors.”) What is the variability over time in concentrations in the media of concern? How has the variability been determined? Has the model been validated? Has the selected model output been compared to those from other similar models, and are they similar (e.g., estimates in the same direction)?
 - Models developed by or accepted by authoritative bodies (e.g., the EPA⁸⁶) are generally considered reliable, at least for the intended applications. Other models should be more carefully evaluated to determine validity and applicability.
- What exposure, over what period of time, by which routes, has been incurred? What calculations support this determination?
- What is the variability among members of the population in their exposure to the chemical of concern? How is this known?
- What is the likely error in the exposure estimates?
- What uncertainties are associated with the exposure/duration findings? Is it a “most likely” estimate, or is it an upper limit? To what fraction of the population is the upper limit likely to apply?
- What has been omitted from the exposure assessment, and why?

These questions are perhaps the minimum that an expert should be able to address when offering testimony. An expert should be able to support such answers with documentation.

86. Such exposure models are enumerated in EPA’s ExpoBox. U.S. Env’t Prot. Agency, EPA ExpoBox (A Toolbox for Exposure Assessors), <https://perma.cc/T6FB-PS3V>.

Qualifications of Exposure Scientists or Other Exposure Assessors

As discussed generally by Liesa L. Richter and Daniel J. Capra in their reference guide in this manual, *The Admissibility of Expert Testimony*,⁸⁷ the court is expected to perform a gatekeeping function with regard to proposed experts involved in the evaluation of exposure-related information and assessment of exposure experts for a given case.

A determination of whether or not a proposed exposure expert is qualified is complicated by the fact that exposure science is a heterogeneous field with various subspecialties (e.g., biomonitoring, environmental monitoring, modeling, remote sensing, use of sensors/dosimeters) as well as specialized applications, such as industrial hygiene/occupational exposures. No single academic degree, research specialty, or career path qualifies an individual as an expert in and of itself. Further complicating evaluation of expert qualifications, a given exposure assessment may involve collaborative efforts among members of various disciplines. However, there are a number of indicia of expertise that can be ascertained for an individual proffering an opinion on exposure, outlined as follows.

Does the Proposed Expert Have an Advanced Degree or Certification in Exposure Science, Industrial Hygiene, Environmental Health, or a Related Field?

Historically, exposure scientists and related experts have possessed a wide range of academic backgrounds, spanning industrial hygiene, environmental and analytical chemistry, chemical and environmental engineering, geology and hydrogeology, toxicology (toxicokinetic applications in particular), epidemiology, and even behavioral sciences (pertaining to those aspects of human behavior that affect exposures). Note that a single course in exposure assessment or risk assessment is unlikely to provide sufficient background for developing expertise in exposure science. An increasing number of schools offer either a graduate degree in the area of exposure science and/or exposure assessment or, more commonly, a concentration or emphasis in these areas within a broader degree program such as Environmental Health Sciences.⁸⁸ Such a focused graduate degree demonstrates

87. See Liesa L. Richter & Daniel J. Capra, *The Admissibility of Expert Testimony*, in this manual.

88. While some institutions have dedicated exposure science degree programs, such as the doctoral program in Human Exposure Assessment at Rutgers University, many leading U.S.

that the proposed expert has a substantial background in the basic issues and tenets of exposure science and exposure assessment. There are currently no certification programs available specifically for exposure scientists, although the Europe Regional Chapter of the International Society of Exposure Science (ISES Europe) has developed a program to begin certifying International Registered Exposure Scientists (IRES) by 2026. However, experts may hold other certifications, many as certified industrial hygienists.⁸⁹

Is the Proposed Expert Established Professionally in the Field?

The success of academic scientists in exposure science, as in other applied sciences, is usually ascertained using the following types of criteria: the quality and number of peer-reviewed publications, the ability to compete for research grants, service on scientific advisory panels and/or working groups, university appointments, and participation or service in professional organizations and societies. Publication of articles in relevant peer-reviewed journals or peer-reviewed technical reports or exposure assessments (as is often more common for governmental experts) generally is construed to indicate some degree of expertise in exposure science. The number of articles, their topics, and whether the individual is the principal or senior author are important factors in determining the expertise of such a scientist.⁹⁰ Selection for local, national, and international regulatory advisory panels or working groups (including those convened by the

institutions with environmental and public health programs now offer various degrees with a concentration or focused track in exposure science and/or exposure assessment, including Harvard University (Environmental Health Exposures/Exposure Assessment specialty), the Johns Hopkins University (Exposure Sciences and Environmental Epidemiology track), the University of Michigan (Exposure Science-Industrial Hygiene track), and the University of Washington (area of emphasis in Occupational Hygiene/Exposure Science).

89. See, e.g., *Allen v. Martin Surfacing*, 263 F.R.D. 47 (D. Mass. 2009) (industrial hygienist qualified to testify regarding concentration and duration of plaintiffs' decedent's exposure to toluene and other chemicals); *Buzzerd v. Flagship Carwash of Port St. Lucie, Inc.*, 669 F. Supp. 2d 514 (M.D. Pa. 2009) (industrial hygienist qualified to opine on carbon monoxide exposure, but his conclusions were not based on reliable methodology).

90. Examples of reputable, peer-reviewed journals that publish exposure-related research are the *Journal of Exposure Science & Environmental Epidemiology*; *American Journal of Industrial Medicine*; *Annals of Work Exposures and Health* (previously *Annals of Occupational Hygiene*); *Archives of Environmental & Occupational Health* (formerly the *Archives of Environmental Health*); *Atmospheric Environment*; *Biomarkers*; *Chemosphere*; *Environmental Health Perspectives*; *Environment International*; *Environmental Research*; *Environmental Science & Technology*; *Environmental Toxicology and Chemistry*; *Exposure and Health*; *Indoor Air*; *Integrated Environmental Assessment and Management*; *International Journal of Environmental Research and Public Health*; *International Journal of Hygiene and Environmental Health*; *Journal of Occupational and Environmental Hygiene*; *Journal of Toxicology and Environmental Health*; *Journal of*

EPA, FDA, National Academies, National Institutes of Health (NIH), and World Health Organization (WHO)) also typically implies recognition in the field. Similarly, exposure-focused research grants from government agencies (including the National Institute of Environmental Health Sciences, the National Institute of Occupational Safety and Health, and the EPA) and private foundations are highly competitive. Successful competition for funding may also be construed to indicate competence in this area. A university appointment in exposure science, industrial hygiene/occupational health, environmental health, risk assessment, or a related field signifies relevant expertise, particularly if the university has a graduate education program in that area and the expert provides formal instruction or mentoring of trainees and fellows.⁹¹ Participation in professional societies⁹² is common for advanced practitioners; not only do meetings and events allow for presentation of research and networking activities, but most societies give awards to notable professionals based on scientific contributions to the field.

Does the Proposed Expert Have Experience Relevant to the Topic or Method of Interest?

Given the heterogeneity of exposure science, a given expert's body of work and training should be carefully evaluated for relevance to the exposure-related issue(s) of interest in a given case. Exposure scientists are increasingly scrutinized not just for the scope of their expertise in general,⁹³ but for their scope of

Occupational and Environmental Medicine; Regulatory Toxicology and Pharmacology; Risk Analysis; Science of the Total Environment; Toxicology; Toxicology Letters; and Toxicological Sciences.

91. See, e.g., *Bearden v. Honeywell Int'l, Inc.*, No. 3:09-CV-1035, 2015 WL 7574344 (M.D. Tenn. Nov. 23, 2015).

92. Note that only one professional society, the International Society of Exposure Science, is dedicated to exposure science and/or exposure assessment. Such membership has been referenced in expert qualifications, see, e.g., *Bearden v. Honeywell Int'l*, *supra* note 91. However, other professional organizations and societies relevant to toxicology and risk assessment have established specialty groups or technical sections in exposure science or assessment, including the Society of Environmental Toxicology and Analytical Chemistry's Exposure Modeling Interest Group, the Society of Toxicology's Exposure Specialty Section, the Society of Risk Analysis' Exposure Assessment Specialty Group, and the American Industrial Hygiene Association's Exposure Assessment Strategies Committee. Of these, only the Society of Toxicology has qualifications-centered requirements for membership, based either on peer-reviewed publications or on the active practice of toxicology or a related discipline that informs toxicology.

93. See, e.g., *United States v. Burkich*, No. 1:19-CV-3510-MLB, 2022 WL 4236243, at *4 (N.D. Ga. Sept. 14, 2022) (exclusion of opinions offered in the area of patient care following exposure by an exposure expert without a background in healthcare based on determination that the expert is "seeking to testify outside the scope of [her] academic and professional specialty") (quoting *Moore v. Intuitive Surgical, Inc.*, 995 F.3d 839, 853 n.12 (11th Cir. 2021)).

expertise within the field itself⁹⁴ as the types of exposure data and exposure issues presented as evidence continue to evolve. For example, a scientist specializing in dispersion modeling of toxic air contaminants may not have sufficient expertise to address exposure and/or dose reconstruction from biomonitoring data. And while some healthcare professionals specializing in occupational and environmental health may have training in certain elements of exposure assessment (particularly those who work in the chemical, petrochemical, and pharmaceutical industries, in which the surveillance of workers exposed to chemicals is a major responsibility), these individuals⁹⁵ may not be sufficiently familiar with exposures outside of occupational settings or with certain aspects of data collection and analysis. While a review of a proposed expert's biosketch, curriculum vitae, and/or publication record should give some indication of areas of specialization, additional focused questions may be necessary.

94. See, e.g., *In re 3M Combat Arms Earplug Prods. Liab. Litig.*, No. 3:19MD2885, 2021 WL 948839 (N.D. Fla. Mar. 13, 2021) (expert's professional experience in the past decade was related to chemical and product exposure and risk assessment, rather than the exposure of interest, which was noise exposure); *Jones v. United States*, No. 2:16-cv-00435-JRS-DLP, 2019 WL 367622 (S.D. Ind. Jan. 30, 2019) (expert's training and professional experience, which were in laser and air particulates, were deemed not relevant to provide an opinion regarding the presence of *H. pylori* in water and the transmission of *H. pylori* through waterborne sources).

95. See David L. Eaton et al., *Reference Guide on Toxicology*, in this manual ("Most practicing physicians have little knowledge of environmental and occupational medicine.").

Appendix A: Units of Exposure and Dose

Choosing the proper units to express concentrations of chemicals in environmental media is crucial for precisely defining exposure. Chemical concentrations in environmental media usually are reported in one of two forms: as numeric ratios, such as parts per million or billion (ppm and ppb, respectively), or as unit weight of the chemical per weight or volume of environmental media, such as milligrams per kilogram (mg/kg) or milligrams per cubic meter (mg/m³). Although concentrations expressed as parts per million or parts per billion are easier for some people to conceptualize, their use assumes that media are always sampled at standard temperature and pressure (25°C and 760 torr, respectively). Consequently, scientists often prefer to express chemical concentrations as weight of chemical per unit weight or volume of media. This method also makes conversions to exposure and dose equivalents, usually expressed in terms of weight of chemical per unit body weight (mg/kg bw), more convenient. To permit the presentation of results without excessive zeroes before or after the decimal point, appropriate units are needed. The choice of units depends on both the medium in which the chemical is present and the amount of chemical measured. For example, if 50 nanograms of chemical were found in one liter of water, the appropriate units would be ng/L, rather than 0.00005 mg/L. If 50 grams were found instead, the appropriate units would be 50,000 mg/L, because milligrams are generally the largest units used to express the mass of a chemical in media (Table 1).

Table 1. Weight of Chemical Per Unit Weight of Medium

Preferred Unit	Alternative Unit
mg/kg	ppm (parts per million)
µg/kg	ppb (parts per billion)
ng/kg	ppt (parts per trillion)
pg/kg	ppq (parts per quadrillion)

In water or food, concentration expressed by the preferred unit equals concentration expressed by alternative unit; thus, 2 mg/kg = 2 ppm. One mg (10⁻³ g) per kg (10³ g) equals 1 part per million (10⁻³/10³ = 10⁻⁶). Similarly, 1 µg (10⁻⁶ g) per kilogram (10³ g) equals 1 part per billion (10⁻⁶/10³ = 10⁻⁹), and so on (Table 2).

Table 2. Weight of Chemical Per Unit Volume of Medium

Water	Air
mg/L = ppm	mg/m ³ ≠ ppm
μg/L = ppb	mg/m ³ ≠ ppb
ng/L = ppt	ng/m ³ ≠ ppt

Note that in air, parts per million and parts per billion have different meanings than they do in water or food; to avoid confusion, it is always preferable to express air concentrations in weight of chemical per unit volume (rather than weight) of air (usually cubic meters, m³).

Glossary of Terms

The following definitions are derived from a range of sources, including the ES21FWG Glossary released in June 2015,⁹⁶ the U.S. EPA *Exposure Factors Handbook*,⁹⁷ various National Academies reports,⁹⁸ and the official glossary adopted by the International Society of Exposure Science.⁹⁹

absorbed dose. The amount of a substance that actually enters the body following absorption.

absorption. The penetration of a substance through a barrier (e.g., the skin, the gut, or the lungs).

absorption barrier. Any exposure surface that may retard the rate of penetration of an agent into an organism. Examples of absorption barriers are the skin, respiratory tract lining, and gastrointestinal tract wall.

accuracy. The ability of a method to determine the “true” concentration of the environment sampled, or the theoretical maximum error of measurement.

activity pattern data. Information on human activities used in exposure assessments. These may include a description of the activity, frequency of activity, duration spent performing the activity, and the microenvironment in which the activity occurs.

acute exposure. A contact between a stressor and an individual occurring over a short time, generally less than a day. (Other terms, such as short-term exposure and single dose, are also used.)

agent. A chemical, biological, or physical stressor that contacts a receptor (individual or population).

aggregate exposure. The sum of an individual’s exposures to a single stressor from all sources, routes, and pathways over a period of time.

analyte. A specific chemical moiety being measured, which can be an intact drug, a biomolecule or its derivative, a metabolite, and/or a degradation product in a biologic matrix.

96. ES21 Fed. Working Grp. on Exposure Sci., Glossary of Exposure Science Terms, <https://perma.cc/8XCY-YPRD>.

97. U.S. Env’t Prot. Agency, *Exposure Factors Handbook* (2011), <https://perma.cc/D8AR-5FT5>. Note: The EPA has developed a companion software tool, ExpoFIRST, that utilizes data from the handbook in development of user-defined exposure scenarios for a given assessment. ExpoFIRST can then be used to calculate deterministic exposure estimates as point estimates for various receptor populations and life stages. This tool is accessible via the EPA website. U.S. Env’t Prot. Agency, EPA ExpoBox (A Toolbox for Exposure Assessors), <https://perma.cc/T6FB-PS3V>.

98. See Nat’l Rsch. Council, *Human Biomonitoring for Environmental Chemicals* (2006), <https://doi.org/10.17226/11700>.

99. Valerie Zartarian, Tina Bahadori & Tom McKone, *Adoption of an official ISEA glossary*, 15 J. Exposure Analysis & Env’t Epidemiology 1 (2005), <https://doi.org/10.1038/sj.jea.7500411>.

average daily dose (ADD). The average dose received on any given day during a period of exposure, expressed in mg/kg body weight per day. Ordinarily used in assessing noncancer risks.

background level. The amount of an agent in a medium (e.g., water, soil) that is not attributed to the source(s) under investigation in an exposure assessment. Background level(s) can be naturally occurring or the result of human activities.

bias. In exposure measurements, the difference between the average measured mass or concentration and reference mass or concentration expressed as a fraction of reference mass or concentration. In epidemiology, this typically refers to any effect at any stage of investigation or inference tending to produce results that depart systematically from the true values.

bioavailability. The rate and extent to which a chemical (or chemical metabolite) enters the general circulation, thereby permitting access to the site of toxic action.

biological matrix. Discrete material of biological origin that can be sampled and processed in a reproducible manner. Examples are blood, serum, plasma, urine, feces, saliva, sputum, breast milk, semen, and various tissues.

biomarker of effect or biomarker of response. A measurable biochemical, physiologic, behavioral, or other alteration in an individual that, depending on the magnitude, can be recognized as associated with an established or possible health impairment or disease.

biomarker of exposure. A chemical, its metabolite, or the product of an interaction between a chemical and some target molecule or cell, that is measured in an individual.

biomarker of susceptibility. A measurable factor, such as genetic polymorphism, nutritional status, or age, that can result in certain individuals being more sensitive to a given exposure.

biomonitoring. A method used to assess human exposure to chemicals by measuring a chemical, its metabolite, or a reaction product in human tissues or specimens, such as blood and urine.

body burden. The total amount of a chemical present or stored in the body. In humans, body burden is an especially important measure of exposure to chemicals that tend to accumulate in fat cells. These chemicals include DDT, PCBs, and dioxins.

bounding estimate. An estimate of exposure, dose, or risk that is higher than that incurred by the person with the highest exposure, dose, or risk in the population being assessed. Bounding estimates are useful in developing statements that exposures, doses, or risks are “not greater than” the estimated value.

chronic exposure. A continuous, recurring, or intermittent long-term contact between a stressor and a receptor population. (Other terms, such as “long-term exposure,” are also used.)

cumulative exposure. The sum of an individual’s exposures to stressors that affect a single biological target over a period of time.

direct exposure. Exposure of a subject who comes into contact with a chemical via the medium in which it was initially released to the environment. Examples include exposures mediated by cosmetics, other consumer products, some food and beverage additives, medical devices, over-the-counter drugs, and single-medium environmental exposures.

dose. The amount of a substance entering a person, usually expressed for chemicals in the form of weight of the substance (generally in milligrams (mg) or micrograms (μg)) per unit of body weight (generally in kilograms (kg)). The time over which it is received must also be specified. The time of interest is typically one day. If the duration of exposure is specified, dose is actually a dose rate and is expressed as mg or $\mu\text{g/kg}$ per day. In human epidemiological studies or experimental animal models, dose typically indicates the amount of chemical or physical agent that is absorbed into the body, not just the amount or concentration to which a person or animal is externally exposed. Some toxicologists may also refer to the external exposure as an external dose.

dose–response assessment. In risk assessment, an analysis of the relationship between the dose administered to a group and the frequency or magnitude of the biological effect (response).

duration of exposure. Toxicologically, there are four general categories describing duration of exposure: acute (one time), subacute (repeated over a short period, generally less than two weeks), subchronic (repeated, generally up to 90 days), and chronic (repeated, for nearly a lifetime).

environmental media. Air, water, soils, and food; consumer products may also be considered media. Chemicals may be directly and intentionally introduced into certain media. Others may move from their sources through one or more media before they reach the media with which people have contact.

exposure. Contact made between a chemical, physical, or biological agent and the outer boundary of an organism. The opportunity to receive a dose through direct contact with a chemical or medium containing a chemical. See also direct exposure; indirect exposure.

exposure assessment. The process of describing, for a population at risk, the amounts of chemicals to which individuals are exposed, or the distribution of exposures within a population, or the average exposure over an entire population. Also a formal step in the risk assessment process.

exposure frequency. The number of times an exposure occurs in a given period; exposure may be continuous, discontinuous but regular (e.g., once daily), or intermittent (e.g., less than daily, with no standard quantitative definition).

exposure pathway. The connected media that transport a chemical from source to receptor populations.

exposure route. The way an agent enters the body after contact (e.g., by ingestion, inhalation, or dermal absorption).

indirect exposure. Often defined as an exposure involving multimedia transport of chemicals from source to exposed individual. Examples include exposures to chemicals deposited onto soils from the air, chemicals released into the ground water beneath a hazardous waste site, or consumption of fruits or vegetables with pesticide residues.

intake. The amount of contact between an organism and a medium containing a chemical; used for estimating the exposure received from a particular medium. In toxicology, this may also refer to the amount of contact between a biological surface with a medium containing a chemical; used for estimating the dose received from a particular medium.

levels. An alternative term for expressing chemical concentration in environmental media. Usually expressed as mass per unit volume or unit weight in the medium of interest.

lifetime average daily dose (LADD). Total dose received over a lifetime multiplied by the fraction of lifetime during which exposure occurs, expressed in mg/kg body weight per day. Ordinarily used for assessing cancer risk.

limit of detection (LOD). Typically, the smallest amount of analyte distinguishable from background, or the lowest concentration of an analyte that the bioanalytical procedure can reliably differentiate from background noise. See, for example, the FDA's Bioanalytical Method Validation Guidance for Industry (<https://www.fda.gov/regulatory-information/search-fda-guidance-documents/bioanalytical-method-validation-guidance-industry>). A common estimate for unbiased analyses, with media blanks not distinguishable from background, is three times the standard error of the calibration graph for low concentrations, divided by the slope (instrument reading per unit mass or per unit concentration of analyte).

limit of quantification (LOQ) or limit of quantitation (LLOQ). Typically refers to the smallest amount of analyte that can be quantitatively determined with acceptable precision and accuracy from background. See, for example, the FDA's Bioanalytical Method Validation Guidance for Industry (<https://www.fda.gov/regulatory-information/search-fda-guidance-documents/bioanalytical-method-validation-guidance-industry>) and the

CDC's NIOSH Manual of Analytical Methods (NMAM), 5th Edition (<https://perma.cc/FU2Q-LNXV>).

metabolite. Chemical compound that results from biotransformation (i.e., metabolism) of a chemical (e.g., phenol in urine is a metabolite of benzene and is representative of benzene absorption in an exposed worker). Note that often a chemical that enters the body is rapidly metabolized or otherwise difficult to measure or distinguish from external contamination. A metabolite may be more stable and also may be eliminated in urine, making it more accessible and easier to measure.

model. Idealized mathematical expression of the relationship between two or more factors (variables). In toxicology, may also describe the use of animals as a substitute for humans in an experimental system to ascertain an outcome of interest.

particulate matter. General term for complex mixtures of solids and aerosols composed of small droplets of liquid, dry solid fragments, and solid cores with liquid coatings that are suspended in the air. These particles vary widely in size, shape, and chemical composition and are typically defined by their diameter for air quality regulatory purposes and human health risk assessment (e.g., PM 10 for particles with a diameter of 10 microns or less and PM 2.5 for particles with a diameter of 2.5 microns or less).

point-of-contact exposure. Exposure expressed as the product of the concentration of the chemical in the medium of exposure and the duration and surface area of contact with the body surface, for example, mg/cm²-hours. Some chemicals do not need to be absorbed into the body but rather produce toxicity directly at the point of contact—for example, the skin, mouth, GI tract, nose, bronchial tubes, or lungs. In such cases, the absorbed dose is not the relevant measure for toxicity; rather, it is the amount of toxic chemical coming directly into contact with the body surface that would be relevant for toxic effects.

population at risk. A group of subjects with the opportunity to be exposed to a chemical.

receptor population. People who could come into contact with a stressor. See exposure pathway.

risk. The nature and probability of occurrence of an unwanted, adverse effect on human life or health or on the environment.

risk assessment. Characterization of the potential adverse effects on human life or health or on the environment. According to the National Research Council's Committee on the Institutional Means for Assessment of Health Risk, human health risk assessment includes the following:

- description of the potential adverse health effects based on an evaluation of results of epidemiologic, clinical, toxicologic, and environmental research (hazard identification);

- extrapolation from those results to predict the type and estimate the extent of health effects in humans under given conditions of exposure (dose–response assessment);
- judgments regarding the number and characteristics of persons exposed at various intensities and durations (exposure assessment);
- summary judgments on the existence and overall magnitude of the public-health problem; and
- characterization of the uncertainties inherent in the process of inferring risk (risk characterization).

setting. The place or situation in which a person is exposed to the chemical. Setting is often modified by the activity a person is undertaking—for example, occupational or in-home exposures.

sensor. A technology that includes small, portable, autonomous, low-cost, real-time devices. Sensors are air or water monitoring technologies supporting very large numbers of measurement locations, including wearable, mobile (e.g., on autonomous/robotic platforms), or stationary applications. Key traits of sensor devices include direct measurement of one or more toxicants, toxins, and/or pathogens; portability; low power draw; turnkey operation; and market price supporting large numbers to be purchased by the public as individuals or community groups.

source. The activity or entity from which the chemical is released for potential human exposure.

stressor. Any entity, stimulus, or condition that can modulate normal functions of the organism or induce an adverse response (e.g., agent, lack of food, drought).

subchronic exposure. Contact between a stressor and an individual that is of intermediate duration between acute and chronic exposures.

subject. An exposed individual, whether a human or an exposed animal or organism in the environment. An exposed individual is sometimes also called a receptor, while a group of exposed individuals is sometimes called a receptor population.

systemic dose. A dose of a chemical within the body—that is, not localized at the point of contact. Thus, skin irritation caused by contact with a chemical is not a systemic effect, but liver damage due to absorption of the chemical through the skin is. Often referred to as target site dose.

total dose. The doses received by more than one route of exposure are added to yield the total dose. See also aggregate exposure.

uncertainty. A limited knowledge of the agreement between data, information, or outcomes relative to an unknown truth. The uncertainty of a measurement is the parameter associated with the result of a measurement

that characterizes the dispersion of the values that could reasonably be attributed to the quantity being measured (the measurand).

validated method. A method that meets or exceeds certain sampling and measurement performance criteria (e.g., Guidelines for Air Sampling and Analytical Method Development and Evaluation (NIOSH Technical Report)).

validated model. A model that meets or exceeds certain criteria, including credible and objective peer review, corroborating the model by evaluating the degree to which it corresponds to the system being modeled, and performing sensitivity and uncertainty analyses (e.g., EPA Guidance on the Development, Evaluation, and Application of Environmental Models).

validation. The confirmation that an observation meets a defined standard or objective reference.

References on Law and Exposure Science

In addition to the in-text legal citations, the following articles may be useful as references.

- William Anderson & Kieran Tuckley, *How Much Is Enough? A Judicial Roadmap to Low Dose Causation Testimony in Asbestos and Tort Litigation*, 42 Am. J. Trial Advoc. 39 (2018).
- Linda S. Birnbaum et al., *Environmental Health Science for Regulatory Decisionmaking*, 21 Duke Env't L. & Pol'y F. 259 (2010).
- Caroline Cecot, *The Data Gap: Promoting Analysis of Exposure-Related Harms*, 69 DePaul L. Rev. 297 (2019).
- Caroline Gillie et al., *Leveraging Science to Inform Proactive and Reactive Risk Management*, 51 Env't L. Rep. 10198 (2021).
- Michael Green, *The Education of the Judiciary: The Sciences Addressing Disease Causation*, 49 Sw. L. Rev. 492 (2020).
- Laura Hall et al., *Litigating Toxic Risks Ahead of Regulation: Biomonitoring Science in the Courtroom*, 31 Stan. Env't L.J. 3 (2012).
- Jeff B. Kray & Sarah J. Wightman, *Contaminants of Emerging Concern: A New Frontier for Hazardous Waste and Drinking Water Regulation*, 32(4) Nat. Res. & Env't 36–40 (2018).
- Albert C. Lin, *Deciphering the Chemical Soup: Using Public Nuisance to Compel Chemical Testing*, 85 Notre Dame L. Rev. 955 (2009).
- Jason B. Miller & Ranjit J. Machado, *Prospective and Retrospective Exposure Assessment*, 22(3) Env't Claims J. 221–29 (2010).
- Megan Noonan, *The Doctor Can't See You Yet: Overcoming the "Injury" Barrier to Medical Monitoring Recovery for PFAS Exposure*, 45 Vt. L. Rev. 287 (2020).
- Catherine A. O'Neill, *Exposed: Asking the Wrong Question in Risk Regulation*, 48 Ariz. St. L.J. 703 (2016).
- Elizabeth V. Young, *Outsourcing the Jury: Barlett v. DuPont and the Role of Alternative Adjudication in Preserving Jury Fairness in Complex Scientific Litigation*, 77 Ohio St. L.J. Furthermore: Sixth Cir. Rev. 15 (2016).

References on Exposure Science

Select Guidelines, Reports, and Other Technical Documents

- National Academies of Sciences, Engineering, and Medicine, *Using 21st Century Science to Improve Risk-Related Evaluations* (2017), <https://doi.org/10.17226/24635>.

- National Research Council, Risk Assessment in the Federal Government: Managing the Process (1983), <https://doi.org/10.17226/366>.
- National Research Council, Human Exposure Assessment for Airborne Pollutants: Advances and Opportunities (1991), <https://doi.org/10.17226/1544>.
- National Research Council, Human Biomonitoring for Environmental Chemicals (2006), <https://doi.org/10.17226/11700>.
- National Research Council, Models in Environmental Regulatory Decision Making (2007), <https://doi.org/10.17226/11972>.
- National Research Council, Exposure Science in the 21st Century: A Vision and a Strategy (2012), <https://doi.org/10.17226/13507>.
- Organisation for Economic Co-operation & Development, Descriptions of Existing Models and Tools Used for Exposure Assessment: Series on Testing and Assessment No. 182 (2012).
- U.S. Centers for Disease Control and Prevention, National Report on Human Exposure to Environmental Chemicals, <https://perma.cc/LS6D-4WWL>.
- U.S. Environmental Protection Agency, Guidelines for Human Exposure Assessment (Nicolle Tulse et al. eds., 2019), <https://perma.cc/WD6E-AGCF>.
- World Health Organization, International Programme on Chemical Safety, Principles of Characterizing and Applying Human Exposure Models (2005), <https://perma.cc/Q36C-3TLZ>.
- World Health Organization, International Programme on Chemical Safety, Uncertainty and Data Quality in Exposure Assessment (2008), <https://perma.cc/Q6L8-X77R>.
- World Health Organization, International Programme on Chemical Safety, WHO Human Health Risk Assessment Toolkit: Chemical Hazards (2d ed. 2012), Geneva, Switzerland: IPCS Harmonization Project, WHO (2021), <https://perma.cc/W9BH-WHDM>.

Select Databases and Tools

- U.S. Centers for Disease Control and Prevention, National Environmental Public Health Tracking Network (Tracking Network), <https://ephrtracking.cdc.gov>.
- U.S. Environmental Protection Agency, CompTox Chemicals, <https://perma.cc/Z3W6-JGPR>.
- U.S. Environmental Protection Agency, ExpoBox (A Toolbox for Exposure Assessors), <https://perma.cc/YNT8-9R3N>.

Select Articles

The following are select articles that reflect advances in exposure science and exposure assessment and highlight potential applications relevant to litigation and risk assessment. The selected articles focus on exposure-related approaches, considerations, and models rather than reports of specific exposure measures, with an emphasis on articles that demonstrate or inform utility of exposure science to risk assessment.

- Kim A. Anderson et al., *Preparation and Performance Features of Wristband Samplers and Considerations for Chemical Exposure Assessment*, 27 J. Exposure Sci. & Env't Epidemiology 551–59 (2017), <https://doi.org/10.1038/jes.2017.9>.
- Jon A. Arnot, *Mass Balance Models for Chemical Fate, Bioaccumulation, Exposure and Risk Assessment*, in *Exposure and Risk Assessment of Chemical Pollution—Contemporary Methodology* 69–91 (Lubomir I. Simeonov & Mahmoud A. Hassanien eds., 2009), <https://doi.org/10.1007/978-90-481-2335-3>
- Jon A. Arnot et al., *Prioritizing Chemicals and Data Requirements for Screening-Level Exposure and Risk Assessment*, 102(11) Env't Health Persp. 1565–70 (2012), <https://doi.org/10.1289/ehp.1205355>.
- Lesia L. Aylward et al., *Variation In Urinary Spot Sample, 24 H Samples, and Longer-Term Average Urinary Concentrations of Short-Lived Environmental Chemicals: Implications For Exposure Assessment and Reverse Dosimetry*, 27 J. Exposure Sci. & Env't Epidemiology 582–90 (2017), <https://doi.org/10.1038/jes.2016.54>.
- Hyunkyung Ban et al., *Impact of Exposure Factor Selection on Deterministic Consumer Exposure Assessment*, 94 Regul. Toxicology & Pharmacology 240–44 (2018), <https://doi.org/10.1016/j.yrtph.2018.02.007>.
- Mark A. Bonnell et al., *Fate and Exposure Modeling in Regulatory Chemical Evaluation: New Directions from Retrospection*, 20(1) Env't Sci. Processes & Impacts 20–31 (2018), <https://doi.org/10.1039/C7EM00510E>.
- Francesca Borghi et al., *Retrospective Exposure Assessment Methods Used In Occupational Human Health Risk Assessment: A Systematic Review*, 17(17) Int'l J. Env't Rsch. & Pub. Health 6190 (2020), <https://doi.org/10.3390/ijerph17176190>.
- Michael S. Breen et al., *A Review Of Air Exchange Rate Models For Air Pollution Exposure Assessments*, 24(6) J. Exposure Sci. & Env't Epidemiology 555–63 (2014), <https://doi.org/10.1038/jes.2013.30>.
- Antonia M. Calafat, *The US National Health and Nutrition Examination Survey and Human Exposure to Environmental Chemicals*, 215(2) Int'l J. Hygiene & Env't Health 99–101 (2012), <https://doi.org/10.1016/j.ijheh.2011.08.014>.
- Antonia M. Calafat, *Contemporary Issues in Exposure Assessment Using Biomonitoring*, 3 Current Epidemiology Rep. 145–53 (2016), <https://doi.org/10.1007/s40471-016-0075-7>.

- Andrew Caplin et al., *Advancing Environmental Exposure Assessment Science to Benefit Society*, 10(1236) *Nature Commc'n* 1 (2019), <https://doi.org/10.1038/s41467-019-09155-4>.
- Elaine A. Cohen Hubal et al., *Exposure Science and the US EPA National Center for Computational Toxicology*, 20 *J. Exposure Sci. & Env't Epidemiology* 231–36 (2010), <https://doi.org/10.1038/jes.2008.70>.
- Susan A. Csiszar et al., *Stochastic Modeling of Near-Field Exposure to Parabens in Personal Care Products*, 27 *J. Exposure Sci. & Env't Epidemiology* 152–59 (2017), <https://doi.org/10.1038/jes.2015.85>.
- Yuxia Cui et al., *Integrating Multiscale Geospatial Environmental Data into Large Population Health Studies: Challenges and Opportunities*, 10(7) *Toxics* 403 (2022), <https://doi.org/10.3390/toxics10070403>.
- Christiaan Delmaar & Joris Meesters, *Modeling Consumer Exposure to Spray Products: An Evaluation of the Consexpo Web and Consexpo Nano Models with Experimental Data*, 30 *J. Exposure Sci. & Env't Epidemiology* 878–87 (2020), <https://doi.org/10.1038/s41370-020-0239-x>.
- Antonio Di Guardo et al., *Environmental Fate and Exposure Models: Advances and Challenges In 21st Century Chemical Risk Assessment*, 20 *Env't Sci. Processes & Impacts* 58–71 (2018), <https://doi.org/10.1039/C7EM00568G>.
- Peter P. Egeghy et al., *Computational Exposure Science: An Emerging Discipline To Support 21st-Century Risk Assessment*, 124(6) *Env't Health Persp.* 697–702 (2016), <https://doi.org/10.1289/ehp.1509748>.
- Peter Fantke et al., *Coupled Near-Field and Far-Field Exposure Assessment Framework for Chemicals in Consumer Products*, 94 *Env't Int'l* 508–18 (2016), <https://doi.org/10.1016/j.envint.2016.06.010>.
- Peter Fantke et al., *Exposure and Toxicity Characterization of Chemical Emissions and Chemicals In Products: Global Recommendations and Implementation In USEtox*, 26 *Int'l J. Life Cycle Assessment* 899–915 (2021), <https://doi.org/10.1007/s11367-021-01889-y>.
- Mark L. Glasgow et al., *Using Smartphones to Collect Time–Activity Data for Long-Term Personal-Level Air Pollution Exposure Assessment*, 26 *J. Exposure Sci. & Env't Epidemiology* 356–64 (2016), <https://doi.org/10.1038/jes.2014.78>.
- W. Greggs et al., *Qualitative Approach to Comparative Exposure in Alternatives Assessment*, 15(6) *Integrated Env't Assessment & Mgmt.* 880–94 (2019), <https://doi.org/10.1002/ieam.4070>.
- Zequin Guo et al., *Recent Advances in Non-Targeted Screening Analysis Using Liquid Chromatography-High Resolution Mass Spectrometry to Explore New Biomarkers for Human Exposure*, 219 *Talanta* 121339 (2020), <https://doi.org/10.1016/j.talanta.2020.121339>.
- Lei Huang, *A Review of Models for Near-Field Exposure Pathways of Chemicals in Consumer Products*, 574 *Sci. Total Env't* 1182–208 (2017), <https://doi.org/10.1016/j.scitotenv.2016.06.118>.

- Kristin K. Isaacs, *Establishing a System of Consumer Product Use Categories to Support Rapid Modeling of Human Exposure*, 30(1) J. Exposure Sci. & Env't Epidemiology 171–83 (2020), <https://doi.org/10.1038/s41370-019-0187-5>.
- Andrew Larkin & Perry Hystad, *Towards Personal Exposures: How Technology Is Changing Air Pollution and Health Research*, 4(4) Current Env't Health Rep. 463–71 (2017), <https://doi.org/10.1007/s40572-017-0163-y>.
- Paul J. Lioy, *Exposure Science: A View of the Past and Milestones For the Future*, 118(8) Env't Health Persp. 1081–90 (2010), <https://doi.org/10.1289/ehp.0901634>.
- Paul J. Lioy & Kirk R. Smith, *A Discussion of Exposure Science in the 21st Century: A Vision and a Strategy*, 121(4) Env't Health Persp. 405–09 (2013), <https://doi.org/10.1289/ehp.1206170>.
- Lidia Morawska et al., *Applications of Low-Cost Sensing Technologies for Air Quality Monitoring and Exposure Assessment: How Far Have They Gone?*, 116 Env't Int'l 286–99 (2018), <https://doi.org/10.1016/j.envint.2018.04.018>.
- Linda Phillips & Jacqueline Moya, *The Evolution of EPA's Exposure Factors Handbook and Its Future as an Exposure Assessment Resource*, 23 J. Exposure Sci. & Env't Epidemiology 13–21 (2013), <https://doi.org/10.1038/jes.2012.77>.
- Linda Phillips et al., *EPA's Exposure Assessment Toolbox (EPA-Expo-Box)*, 25(2) J. Env't Informatics 81–84 (2015), <https://doi.org/10.3808/jei.2014.00269>.
- Stephen M. Rappaport, *Implications of the Exposome for Exposure Science*, 21 J. Exposure Sci. & Env't Epidemiology 5–9 (2011), <https://doi.org/10.1038/jes.2010.50>.
- Aduldatch Sailabaht et al., *Extension of the Advanced REACH Tool (ART) to Include Welding Fume Exposure*, 15(10) me 2199 (2018), <https://doi.org/10.3390/ijerph15102199>.
- Samantha M. Samon et al., *Silicone Wristbands As Personal Passive Sampling Devices: Current Knowledge, Recommendations for Use, and Future Directions*, 169 Env't Int'l 107339 (2022), <https://doi.org/10.1016/j.envint.2022.107339>.
- Paul T. Scheepers et al., *Application of Biological Monitoring for Exposure Assessment Following Chemical Incidents: A Procedure for Decision-Making*, 21(3) J. Exposure Sci. & Env't Epidemiology 247–61 (2011), <https://doi.org/10.1038/jes.2010.4>.
- Linda S. Sheldon & Elaine A. Cohen Hubal, *Exposure as Part of a Systems Approach for Assessing Risk*, 117(8) Env't Health Persp. 1181–94 (2009), <https://doi.org/10.1289/ehp.0800407>.
- Fenna C. Sillé et al., *The Exposome: A New Approach for Risk Assessment*, 37(1) ALTEX: Alt. Animal Experimentation 3–23 (2020), <https://doi.org/10.14573/altex.2001051>.

- Jon R. Sobus et al., *Integrating Tools for Non-Targeted Analysis Research and Chemical Safety Evaluations at the US EPA*, 28 J. Exposure Sci. & Env't Epidemiology 411–26 (2018), <https://doi.org/10.1038/s41370-017-0012-y>.
- Andrea Spinazzè et al., *How to Obtain a Reliable Estimate of Occupational Exposure? Review and Discussion of Models' Reliability*, 16(15) Int'l J. Env't Rsch. Public Health 2764 (2019), <https://doi.org/10.3390/ijerph16152764>.
- Asimina Stamatielopoulou et al., *Assessing and Enhancing the Utility of Low-Cost Activity and Location Sensors for Exposure Studies*, 190(3) Env't Monitoring & Assessment 155 (2018), <https://doi.org/10.1007/s10661-018-6537-2>.
- Lindsay W. Stanek et al., *Environmental public health research at the US Environmental Protection Agency: A blueprint for exposure science in a connected world*, J. Exposure Sci. & Env't Epidemiology (2024), <https://doi.org/10.1038/s41370-024-00720-8>.
- Susanne Steinle et al., *Quantifying Human Exposure to Air Pollution—Moving from Static Monitoring to Spatio-Temporally Resolved Personal Exposure Assessment*, 443 Sci. Total Env't 184–93 (2013), <https://doi.org/10.1016/j.scitotenv.2012.10.098>.
- Yu-Mei Tan et al., *Aggregate Exposure Pathways In Support of Risk Assessment*, 9 Current Op. in Toxicology 8–13 (2018), <https://doi.org/10.1016/j.cotox.2018.03.006>.
- Susan Viegas et al., *Biomonitoring as an Underused Exposure Assessment Tool in Occupational Safety and Health Context—Challenges and Way Forward*, 17(16) Int'l J. Env't Rsch. & Public Health 5884 (2020), <https://doi.org/10.3390/ijerph17165884>.
- Lance A. Wallace et al., *Validation of Continuous Particle Monitors for Personal, Indoor, and Outdoor Exposures*, 21 J. Exposure Sci. & Env't Epidemiology 49–64 (2011), <https://doi.org/10.1038/jes.2010.15>.
- John F. Wambaugh et al., *New Approach Methodologies for Exposure Science*, 15 Current Op. Toxicology 76–92 (2019), <https://doi.org/10.1016/j.cotox.2019.07.001>.
- EunHye Yoo et al., *Geospatial Estimation of Individual Exposure to Air Pollutants: Moving from Static Monitoring to Activity-Based Dynamic Exposure Assessment*, 105(5) Annals of the Ass'n Am. Geographers 915–26 (2015), <http://www.jstor.org/stable/24537962>.
- Bruce M. Young et al., *Comparison of four probabilistic models (CARES, Calendex, ConsExpo, and SHEDS) to estimate aggregate residential exposures to pesticides*, 22 J. Exposure Sci. & Env't Epidemiology 522–32 (2012), <https://doi.org/10.1038/jes.2012.54>.
- Valerie Zartarian et al., *Adoption of an Official ISEA Glossary*, 15 J. Exposure Sci. & Env't Epidemiology 1–5 (2005), <https://doi.org/10.1038/sj.jea.7500411>.

- Valerie G. Zartarian & Bradley D. Schultz, *The EPA's Human Exposure Research Program for Assessing Cumulative Risk in Communities*, 20 J. Exposure Sci. & Env't Epidemiology 351–58 (2010), <https://doi.org/10.1038/jes.2009.20>.
- Christopher Zuidema et al., *Estimating Personal Exposures from a Multi-Hazard Sensor Network*, 30 J. Exposure Sci. & Env't Epidemiology 1013–22 (2020), <https://doi.org/10.1038/s41370-019-0146-1>.

Reference Guide on Epidemiology

STEVE C. GOLD, MICHAEL D. GREEN, JONATHAN CHEVRIER,
AND BRENDA ESKENAZI

Steve C. Gold, J.D., is Professor of Law and Judge Raymond J. Dearie Scholar, Rutgers Law School, Rutgers University–Newark, Rutgers, The State University of New Jersey.

Michael D. Green, J.D., is Visiting Professor, Washington University in St. Louis School of Law.

Jonathan Chevrier, Ph.D., is Associate Professor of Epidemiology, Department of Epidemiology, Biostatistics and Occupational Health, School of Population and Global Health, Faculty of Medicine and Health Sciences, McGill University.

Brenda Eskenazi, Ph.D., is Professor Emeritus of Maternal and Child Health and Epidemiology, School of Public Health, University of California at Berkeley.

CONTENTS

Introduction, 900

The Different Kinds of Epidemiologic Studies, 905

Experimental and Observational Studies of Suspected Toxic Agents, 905

Types of Observational Study Design, 907

Cohort Studies, 908

Case-Control Studies, 910

Cross-Sectional Studies, 911

Ecological Studies, 912

Genetic and Molecular Epidemiologic Studies, 914

Epidemiologic and Toxicologic Studies, 918

Interpreting the Results of an Epidemiologic Study, 921

Relative Risk, 923

Odds Ratio, 925

Attributable Risk, 927

Internal and External Validity, 928

Sources of Error in Epidemiologic Studies, 929

Sampling Error, 931

False Positives and Statistical Significance, 932

False Negatives, 940

Statistical Power, 941

- Biases, 942
 - Selection Bias, 943
 - Information Bias, 945
 - Exposure information bias, 946
 - Disease information bias, 951
 - Information bias due to misclassification, 953
 - Confounding Bias, 954
 - Techniques to identify confounding factors, 958
 - Techniques to prevent or limit confounding, 959
 - Techniques to control for confounding factors, 960
 - Conceptual Error, 964
 - Biases That Affect a Body of Epidemiologic Evidence, 965
 - Effect Modification, 966
- General Causation, 971
 - Temporal Relationship, 975
 - Strength of the Association, 977
 - Dose–Response Relationship, 977
 - Replication of the Findings, 981
 - Biological Plausibility, 982
 - Consideration of Alternative Explanations, 984
 - Cessation of Exposure, 984
 - Specificity of the Association, 984
 - Consistency with Other Relevant Knowledge, 985
 - Background Rate of Disease, 985
- Methods for Synthesizing or Combining the Results of Multiple Studies, 986
- Specific Causation, 989
- Qualifications of Experts in Epidemiology, 1003
- Acknowledgments, 1007
- Glossary of Terms, 1008
- References, 1023
 - On Epidemiology, 1023
 - On Law and Epidemiology, 1023

FIGURES

1. Design of a cohort study, 908
2. Design of a case-control study, 910
3. Risks in exposed and unexposed groups, 927
4. Hypothetical probability distribution under the null hypothesis for an estimated relative risk of 4.0, 936
5. 90% and 95% confidence intervals for a hypothetical study with a relative risk of 1.5, 939

Reference Guide on Epidemiology

6. Directed acyclic graph (DAG) for a hypothetical study of glyphosate and non-Hodgkin lymphoma, 958
7. Directed acyclic graph for a hypothetical study of air pollution and heart disease, 959
8. Age-adjusted lung-cancer mortality rates per 100,000 person-years by urban-rural classification and smoking category, 961
9. Linear dose response, 978
10. Linear threshold dose response, 978
11. Supra-linear dose response, 980
12. Sub-linear dose response, 980
13. Non-monotonic dose response, 981

TABLES

1. Cross Tabulation of Exposure by Disease Status, 923
2. Cross Tabulation of Cases and Controls by Exposure Status, 926
3. Case-Control Study Outcome, 926
4. Hypothetical Emphysema Study Data, 957

Introduction

Epidemiology is the field of public health and medicine that studies the occurrence, distribution, and determinants of health and disease in human populations. The purpose of epidemiology is to better understand disease causation and to prevent disease in groups of individuals.¹ Epidemiology assumes that disease is not distributed randomly in a group of individuals and that identifiable subgroups, including those exposed to certain agents, are at increased risk of contracting particular diseases.²

Judges and juries increasingly are presented with epidemiologic evidence as the basis of an expert's opinion on causation.³ In the courtroom, epidemiologic research findings are offered to establish or dispute whether exposure to an agent⁴ caused a harmful effect or disease.⁵ Since *Daubert v. Merrell Dow*

1. More generally, epidemiologists study and seek to improve health outcomes in populations; this may go beyond the study of “diseases.” For example, an epidemiologist might be interested in suspected causes of osteoarthritis of the knee (a health outcome that is a disease) and also in interventions thought to improve osteoarthritic knees’ range of motion (a health outcome that is not a disease). In this reference guide, we generally refer to the dependent variable being assessed in an epidemiologic study as a *health outcome*. In certain contexts, however, we use the term *disease*, because a plaintiff’s disease is the type of outcome most often at issue in cases in which epidemiology serves as evidence of causation. In any case, in the legal context, references to health outcomes or diseases should be understood, unless stated otherwise, as constituting adverse changes that are legally cognizable. In addition, in this reference guide we generally discuss outcomes as if they are discrete variables that either are present or absent (such as cancer), even though many health outcomes are continuous variables (such as blood pressure or body mass index).

2. Epidemiologists also conduct studies of beneficial agents that reduce disease risk or that prevent or cure disease or otherwise improve health outcomes.

3. Epidemiologic studies have been well received by courts deciding cases involving toxic substances. See, e.g., *Norris v. Baxter Healthcare Corp.*, 397 F.3d 878, 882 (10th Cir. 2005) (“[E]pidemiology is the best evidence of general causation in a toxic tort case.”); *Newman ex rel. Newman v. McNeil Consumer Healthcare*, No. 10 C 1541, 2013 WL 9936293, at *10 (N.D. Ill. Mar. 29, 2013) (“Epidemiological studies are important . . . [and provide] ‘the primary, generally accepted methodology for demonstrating a causal relation between a chemical and a set of symptoms or a disease.’” (quoting *Conde v. Velsicol Chem. Corp.*, 804 F. Supp. 972, 1025–26 (S.D. Ohio 1992), *aff’d*, 24 F.3d 809 (6th Cir. 1994))). Well-conducted studies are uniformly admitted. 3 *Modern Scientific Evidence: The Law and Science of Expert Testimony* § 23.1, at 321 (David L. Faigman et al. eds., 2021–2022) [hereinafter *Modern Scientific Evidence*].

4. We use the term *agent* to refer to any substance external to the human body that potentially causes disease or other health effects. Drugs, medical devices, chemicals, radiation, vaccines, pathogens (e.g., viruses or bacteria), and minerals (e.g., asbestos) are all agents whose toxicity an epidemiologist might explore. A single agent or a number of independent agents may cause disease, or the combined presence of two or more agents may be necessary for the development of the disease. Epidemiologists also conduct studies of individual characteristics that might pose risks, such as blood pressure and diet, but those studies are rarely of interest in judicial proceedings except as possible alternative causes to an alleged tortious cause. Epidemiologists also may conduct studies of drugs and other pharmaceutical products to assess their efficacy and safety in clinical trials.

5. E.g., *In re Testosterone Replacement Therapy Prods. Liab. Litig.* Coordinated Pretrial Proc., No. 14 C 1748, 2017 WL 1833173 (N.D. Ill. May 8, 2017) (assessing whether plaintiffs’ arterial

Pharmaceuticals,⁶ the predominant use of epidemiologic studies is in connection with motions to exclude the testimony of expert witnesses. Courts deciding such motions routinely address epidemiologic studies and whether they are sufficient to support an expert's causation testimony.⁷

Epidemiology aims to develop evidence to identify agents that are associated with an increased risk of disease in groups of individuals, quantify the amount of excess disease that is associated with an agent, and attempt to provide a profile of the type of individual who is more likely to contract a disease after being exposed to an agent. Epidemiology focuses on the question of general causation (i.e., is the agent capable of causing disease?) rather than that of specific causation (i.e., did it cause disease in a particular individual?).⁸ For example, in the 1950s, Doll and Hill and others published articles about the increased risk of lung cancer in cigarette smokers. Doll and Hill's studies showed that smokers

cardiovascular injuries or venous thromboembolisms were caused by prescription testosterone-replacement-therapy drugs); *In re Mirena IUD Prods. Liab. Litig.*, 169 F. Supp. 3d 396 (S.D.N.Y. 2016) (assessing whether uterine perforation was caused by intrauterine devices); *In re Lipitor (Atorvastatin Calcium) Mktg., Sales Pracs. & Prods. Liab. Litig.*, 174 F. Supp. 3d 911 (D.S.C. 2016) (assessing whether Lipitor caused the development of type 2 diabetes); *In re E.I. du Pont de Nemours & Co. C-8 Pers. Inj. Litig.*, No. 2:13-MD-2433, 2015 WL 4092866 (S.D. Ohio July 6, 2015) (assessing whether water sources contaminated with ammonium perfluorooctanoate (C-8, or PFOA) caused residents to develop certain diseases, including testicular cancer and preeclampsia).

6. 509 U.S. 579 (1993).

7. See Michael D. Green & Joseph Sanders, *Admissibility Versus Sufficiency: Controlling the Quality of Expert Witness Testimony*, 50 Wake Forest L. Rev. 1057 (2015). Often the expert witness in a case in which a study bears on causation is not the investigator who conducted the study. See, e.g., *DeLuca v. Merrell Dow Pharms., Inc.*, 911 F.2d 941, 953 (3d Cir. 1990) (holding a pediatric pharmacologist expert's credentials sufficient pursuant to Fed. R. Evid. 702 to interpret epidemiologic studies and render an opinion based thereon); *In re Deepwater Horizon Belo Cases*, No. 3:19CV963-MCR-HTC, 2022 WL 17721595, at *13 (N.D. Fla. Dec. 15, 2022) (stating that non-practicing physician who worked in public health and biomedical research was qualified to testify about general causation); *Donner v. Alcoa Inc.*, No. 10-CV-00908-DW, 2014 WL 12600281, at *5 (W.D. Mo. Dec. 19, 2014) (holding admissible a pathologist's opinion about general and specific causation regarding exposure to aluminum dust and pulmonary fibrosis); *Watson v. Dillon Cos., Inc.*, 797 F. Supp. 2d 1138, 1162 (D. Colo. 2011) (holding medical doctors' opinions about general and specific causation regarding inhalation of butter flavoring ingredients in microwave popcorn and a rare lung condition were admissible); *Sugarman v. Liles*, 190 A.3d 344, 345–68 (Md. 2018) (holding admissible a physician's testimony about general and specific causation regarding lead paint and cognitive injuries).

8. The distinction between general causation and specific causation is widely recognized in court opinions. See, e.g., *Norris v. Baxter Healthcare Corp.*, 397 F.3d 878, 881 (10th Cir. 2005) ("Plaintiff[s] must first demonstrate general causation because without general causation, there can be no specific causation."); *Harrison v. BP Expl. & Prod. Inc.*, No. CV 17-4346, 2022 WL 2390733, at *4 (E.D. La. July 1, 2022) ("Once a plaintiff's diagnoses have been confirmed, the plaintiff has the burden of establishing general causation and specific causation."); *Rhynne v. United States Steel Corp.*, 474 F. Supp. 3d 733, 743 (W.D.N.C. 2020) ("Plaintiffs must prove both general causation and specific causation."). For a discussion of specific causation, see the section titled "Specific Causation" below.

who smoked 10 to 20 cigarettes a day had a lung-cancer mortality rate that was about 10 times higher than that of nonsmokers.⁹ These studies identified an association between smoking cigarettes and death from lung cancer that contributed to the determination that smoking causes lung cancer.

However, it should be emphasized that an association is not equivalent to causation.¹⁰ An association is the relationship between two events (e.g., exposure to a chemical agent and development of disease) that occur more frequently together than one would expect by chance. An association identified in an epidemiologic study may or may not be causal.¹¹

Assessing whether an association is causal requires an understanding of the strengths and weaknesses of the study's design and implementation, as well as a judgment about how the study findings fit with other similar studies and

9. Richard Doll & A. Bradford Hill, *Lung Cancer and Other Causes of Death in Relation to Smoking: A Second Report on the Mortality of British Doctors*, 2 Brit. Med. J. 1071 (1956), <https://doi.org/10.1136/bmj.2.5001.1071>.

10. In *In re Lipitor* (Atorvastatin Calcium) Mktg., Sales Pracs. & Prod. Liab. Litig., 227 F. Supp. 3d 452 (D.S.C. 2017), *aff'd sub nom. In re Lipitor* (Atorvastatin Calcium) Mktg., Sales Pracs. & Prod. Liab. Litig. (No II) MDL 2502, 892 F.3d 624 (4th Cir. 2018), the court explained the relationship between an association and causation:

Establishing an association is the first threshold step in establishing general causation, and it is not surprising that courts may invoke this language to help differentiate the inquiries of general and specific causation. However, this fact does not change voluminous and well-established precedent that association, alone, is not sufficient to establish causation and does not change the simple fact that association is not causation. The parties have always agreed that establishing association is just the first step of a two-step process for establishing general causation.

Id. at 483 n.23. *See also* Harris v. CSX Transp., Inc., 753 S.E.2d 275, 282 (W. Va. 2013) (“It should be clearly understood that the term “association” is a term of art in epidemiology . . . [and] is not the same as causation. An epidemiological association identified in a study may or may not be causal.”); Soldo v. Sandoz Pharms. Corp., 244 F. Supp. 2d 434, 461 (W.D. Pa. 2003) (discussing Hill criteria developed to assess whether an association is causal; *see* section titled “General Causation” below); Magistrini v. One Hour Martinizing Dry Cleaning, 180 F. Supp. 2d 584, 591 (D.N.J. 2002) (“[A]n association is not equivalent to causation.” (quoting the second edition of this reference guide). Association is more fully discussed in the section titled “Interpreting the Results of an Epidemiologic Study” below.

Causation is used to describe the relation between two events when one event (the cause) is a necessary link in a chain of events that results in the effect. Of course, alternative causal chains may exist that do not include the agent but that result in the same effect. For general treatment of causation in tort law and an explanation that for factual causation to exist an agent must be a necessary link in a causal chain sufficient for the outcome, *see Restatement (Third) of Torts: Liability for Physical and Emotional Harm* § 26 (2010). Epidemiologic methods cannot deductively prove causation; indeed, all empirically based science cannot directly prove a causal relation. *See, e.g.,* Kenneth J. Rothman & Sander Greenland, *Causation and Causal Inference in Epidemiology*, 95 Am. J. Pub. Health S144 (2005), <https://doi.org/10.2105/AJPH.2004.059204>. However, epidemiologic evidence can justify an inference, and sometimes a very strong inference, that an agent causes a disease. *See* section titled “Effect Modification” below.

11. *See* section titled “Sources of Error in Epidemiologic Studies” below.

scientific knowledge. It is important to emphasize that all studies have limitations that must be considered to interpret their results properly.¹² Some limitations are inevitable given the limits of technology, resources, the ability and willingness of persons to participate in a study, participant burden, and ethical constraints. In evaluating epidemiologic evidence, the key questions, then, are the extent to which a study's limitations compromise its findings and permit inferences about causation.

Judges should also appreciate that, with some frequency, courts may confront cases in which there is little or no epidemiologic evidence available that addresses the agent–disease causation issue in dispute. Epidemiology studies can be expensive and time-consuming to conduct. Other constraints, such as the rarity of exposure to the suspected toxicant, may prevent development of meaningful epidemiologic evidence. Cases in which agent–disease causation is strongly disputed also tend to be ones in which there is not a robust body of epidemiology that provides persuasive evidence of the existence of general causation *vel non*.

Epidemiology studies risk in samples of populations. Employing the results of group-based studies of risk to make a causal determination for an *individual* plaintiff is beyond the limits of epidemiology. Nevertheless, a substantial body of legal precedent has developed that addresses the use of epidemiologic evidence to prove causation for an individual litigant through probabilistic means. The law developed in these cases is discussed later in this reference guide.¹³

The following sections of this reference guide address a number of critical issues that arise in considering the admissibility of, and weight to be accorded to, epidemiologic research findings. A glossary of oft-encountered epidemiologic terms is appended to the guide.¹⁴

Over the past several decades, courts frequently have confronted the use of epidemiologic studies as evidence and have recognized their utility in proving causation. As the Third Circuit observed in *DeLuca v. Merrell Dow Pharmaceuticals*: “The reliability of expert testimony founded on reasoning from epidemiologic

12. See *In re Phenylpropanolamine (PPA) Prods. Liab. Litig.*, 289 F. Supp. 2d 1230, 1240 (W.D. Wash. 2003) (quoting the second edition of this reference guide and criticizing defendant’s “ex post facto dissection” of a study, recognizing that “scientific studies almost invariably contain flaws.”); *Henricksen v. ConocoPhillips Co.*, 605 F. Supp. 2d 1142, 1169 (E.D. Wash. 2009) “[T]his court understands that in epidemiology hardly any study is ever conclusive, and the court does not suggest that an expert must back his or her opinion with published studies that unequivocally support his or her conclusions.”); *Barrera v. Monsanto Co.*, 2019 WL 2331090, at *8 (Del. Super. Ct. May 31, 2019) (discussing flaws in cohort and case control studies relied on by expert witnesses); Joseph L. Gastwirth, *Reference Guide on Survey Research*, 36 *Jurimetrics J.* 181, 185 (1996), <https://www.jstor.org/stable/29762414> (review essay) (“One can always point to a potential flaw in a statistical analysis.”).

13. See section titled “Specific Causation” below.

14. See Glossary of Terms below.

data is generally a fit subject for judicial notice; epidemiology is a well-established branch of science and medicine, and epidemiologic evidence has been accepted in numerous cases.”¹⁵ Indeed, much more difficult problems arise for courts when there is a paucity of epidemiologic evidence.

Three basic issues arise when epidemiology is used in legal disputes and the methodological soundness of a study and its implications for resolution of the question of causation must be assessed:

1. Do the results of an epidemiologic study or studies reveal an association between an agent and disease?
2. Could this association have resulted from limitations of the study (bias or sampling error), and if so, from which?
3. Based on the analysis of limitations in Question 2 above, and on other evidence, how plausible is a causal interpretation of the association?

In this reference guide, the section titled “The Different Kinds of Epidemiologic Studies” explains the nature and relative strengths and weaknesses of various types of epidemiologic research designs; the “Interpreting the Results of an Epidemiologic Study” section addresses the meaning of their outcomes, including discussion of how to assess the validity of an epidemiologic study and of how to evaluate the existence and significance of several potential sources of error.¹⁶ The “General Causation” section discusses general causation, considering whether an agent is capable of causing disease. The “Methods for Synthesizing or Combining the Results of Multiple Studies” section deals with methods for combining the results of multiple epidemiologic studies and the difficulties entailed in extracting a single global measure of risk from multiple studies. The “Specific Causation” section addresses the matter of whether a specific agent caused the disease in a given plaintiff, including the recurrent issue of whether

15. 911 F.2d 941, 954 (3d Cir. 1990); *see also* *Norris v. Baxter Healthcare Corp.*, 397 F.3d 878, 881, 882 (10th Cir. 2005) (“We agree with the district court that epidemiology is the best evidence of general causation in a toxic tort case.”); *Riddell-Hare v. BP Expl. & Prod., Inc.*, No. CV 17-4177, 2022 WL 3445718, at *4 (E.D. La. Aug. 17, 2022) (observing that the “Fifth Circuit has held that epidemiology provides the best evidence of causation in a toxic tort case”); *Brasher v. Sandoz Pharms. Corp.*, 160 F. Supp. 2d 1291, 1296 (N.D. Ala. 2001) (“Unquestionably, epidemiologic studies provide the best proof of the general association of a particular substance with particular effects, but it is not the only scientific basis on which those effects can be predicted.”); *In re Accutane Litig.*, 191 A.3d 560, 576 (N.J. 2018) (explaining expert’s testimony that epidemiologic studies provided the best available evidence on the disputed causal issue).

16. For a more in-depth discussion of the statistical basis of epidemiology, *see* David H. Kaye & Hal S. Stern, *Reference Guide on Statistics and Research Methods*, in this manual, and two case studies: Joseph Sanders, *The Bendectin Litigation: A Case Study in the Life Cycle of Mass Torts*, 43 *Hastings L.J.* 301 (1992), <https://perma.cc/C2LN-SB53>; Devra L. Davis et al., *Assessing the Power and Quality of Epidemiologic Studies of Asbestos-Exposed Populations*, 1 *Toxicological & Indus. Health* 93 (1985), <https://doi.org/10.1177/074823378500100407>; *see also* section titled “References” below.

and how population-based epidemiologic evidence can be used to address specific causation *vel non*.

The Different Kinds of Epidemiologic Studies

Experimental and Observational Studies of Suspected Toxic Agents

To determine whether an agent affects the risk of developing a certain disease or an adverse health outcome, we might ideally want to conduct an experimental study in which the subjects would be randomly assigned to one of two groups: one group exposed to the agent of interest and the other not exposed. After a period of time, the study participants in both groups would be evaluated for the development of the disease. This type of study, called a randomized trial, randomized controlled trial (RCT), clinical trial, or true experiment, is considered the gold standard to estimate the causal effect of an agent on a health outcome or adverse side effect. Such a study design is often used to evaluate new drugs or medical treatments and is the best way to determine whether the observed difference in outcomes between the two groups is caused by exposure to the drug or medical treatment.

Randomization minimizes the likelihood that there are differences in relevant characteristics between those exposed to the agent and those not exposed. Researchers conducting clinical trials generally attempt to use study designs that are placebo-controlled, which means that the group not receiving the active agent or treatment is given an inactive ingredient that appears similar to the active agent under study. Where possible, they are also double-blinded, which means that neither the participants nor those conducting the study know which group is receiving the agent or treatment and which group is receiving the placebo. However, ethical and practical constraints limit the use of such experimental methodologies to studies that aim to assess the effects of agents or interventions that are potentially beneficial to human beings or the withdrawal of an agent thought to be harmful, such as studies of smoking cessation.¹⁷

17. Although clinical trials cannot intentionally expose subjects to suspected toxicants, those studies can provide evidence that a new drug or other beneficial intervention also has adverse effects. See *In re Lipitor* (Atorvastatin Calcium) Mktg., Sales Pracs. & Prods. Liab. Litig., 227 F. Supp. 3d 452, 481 & n.19 (D.S.C. 2017) (explaining clinical trial that found an association between Lipitor and diabetes), *aff'd*, 892 F.3d 624 (4th Cir. 2018); *In re Bextra & Celebrex* Mktg. Sales Pracs. & Prod. Liab. Litig., 524 F. Supp. 2d 1166, 1181 (N.D. Cal. 2007) (relying on a clinical study of Celebrex that revealed increased cardiovascular risk to conclude that the plaintiff's experts' testimony on causation was admissible).

When an agent's effects are suspected to be harmful, researchers cannot knowingly expose people to the agent.¹⁸ Instead, epidemiologic studies typically "observe" a group of individuals who have been exposed to an agent of interest, such as cigarette smoke or an industrial chemical, and compare them with another group of individuals who have not been exposed.¹⁹ Thus, the investigator identifies a group of subjects who have been exposed²⁰ and compares their rate of disease or death with that of an unexposed group or compares health outcomes among groups of individuals with different levels of exposure.

In contrast to clinical trials, in which other potential risk factors can be controlled by randomizing exposure, observational epidemiologic studies entail the possibility that some other characteristic (a confounder)²¹ may be nonrandomly distributed between the exposed and unexposed groups, which could distort a study's results.²²

Epidemiologic investigators address the possible role of these other characteristics—such as sex, age, social class, diet, exercise, exposure to other environmental agents, and genetic background—by measuring them when possible and by considering them in the design of the study and in the analysis and

18. See, e.g., Gordon H. Guyatt, *Using Randomized Trials in Pharmacoepidemiology*, in Drug Epidemiology and Post-Marketing Surveillance 59 (Brian L. Strom & Giampaolo Velo eds., 1992), https://doi.org/10.1007/978-1-4899-2587-9_8; Comm'n on Use of Third Party Toxicity Rsch. with Hum. Rsch. Participants, Nat'l Rsch. Council, Intentional Human Dosing Studies for EPA Regulatory Purposes: Scientific and Ethical Issues 9 (2004), <https://doi.org/10.17226/10927> ("No study is ethically justifiable if it is expected to cause lasting harm to study participants."); 45 C.F.R. § 46.111 (2022) (requiring that in federally funded research on human subjects, risks to subjects be minimized and be reasonable in relation to anticipated benefits to subjects and the importance of knowledge gained); see also *McClellan v. I-Flow Corp.*, 710 F. Supp. 2d 1092, 1109 (D. Or. 2010) ("ethical considerations preclude randomized, controlled epidemiological studies of continuous infusion given the [risk of harm]"). Experimental studies can be used where the agent under investigation is believed to be beneficial, as is the case in the development and testing of new pharmaceutical drugs. See, e.g., *McDarby v. Merck & Co.*, 949 A.2d 223, 270 (N.J. Super. Ct. App. Div. 2008) (involving an expert witness relying on a clinical trial of a new drug to find the adjusted risk for the plaintiff).

19. Classifying these studies as observational in contrast to randomized trials can be misleading to those who are unfamiliar with the area, because subjects in a randomized trial are observed as well. Nevertheless, the term observational studies is widely used to distinguish them from experimental studies.

20. The subjects may have voluntarily exposed themselves to the agent of interest, as is the case, for example, for those who smoke cigarettes, or subjects may have been exposed to an agent involuntarily or even without their knowledge, such as in the case of employees who are exposed to chemical fumes at work.

21. Confounding is a type of bias, discussed in the section titled "Biases" below.

22. Both experimental and observational studies are subject to random error. See section titled "Sampling Error" below.

interpretation of the study results (see section titled “Sources of Error in Epidemiologic Studies” below).²³

Types of Observational Study Design

Several different types of observational epidemiologic studies can be conducted. Study designs may be selected based on their suitability to investigate the question of interest, feasibility, timing and ethical constraints, resource limitations, or other considerations.

Most observational studies collect data about both exposure and health outcome in every individual in the study. While many study designs exist, the three main types of observational studies are cohort, case-control, and cross-sectional studies. These studies collect data about a sample of individuals selected from a “source” population, from which the researcher seeks to make inferences about the population.²⁴ A final type of observational study is an ecological study, which uses aggregate data collected about groups of people rather than individuals.²⁵

The difference between cohort and case-control studies is that cohort studies measure and compare health outcomes (such as the presence or absence of a particular disease)²⁶ in the exposed and unexposed groups (or between individuals that experienced different levels of exposure), while case-control studies measure and compare the frequency (or level) of exposure in the group with the disease (the cases) and the group without the disease (the controls). In a cohort study, the subjects’ exposure is determined before their health status (see Figure 1). The risk of disease among the exposed can then be compared with the risk of disease among the unexposed. In a case-control study, the health status is determined first. The odds that someone with the disease was exposed to a suspected agent can then be compared with the odds that someone without the disease was similarly exposed. As for most epidemiologic studies, these study designs aim to determine if there is an association between exposure to an agent and a health outcome and the strength (magnitude) of that association.

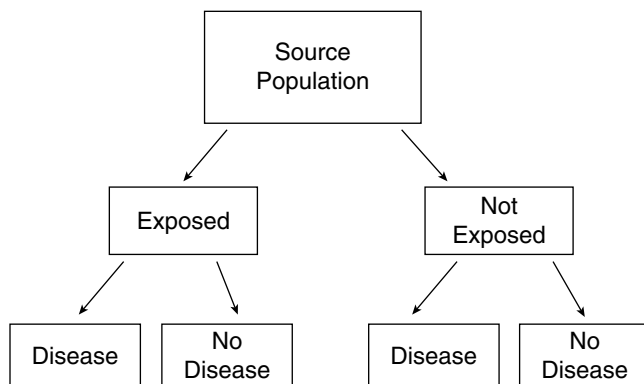
23. See David A. Freedman, *Editorial: Oasis or Mirage?*, 21 *Chance* no. 1, 59–61 (2008), <https://doi.org/10.1080/09332480.2008.10722888>.

24. Sometimes, for practical reasons, the source population from which the researcher draws study subjects is not exactly the same as the population about which the researcher wants to make inferences, in which case the term *target population* is used to distinguish the population about which inferences are made from the source population.

25. See section titled “Ecological Studies” below. For thumbnail sketches on all types of epidemiologic study designs, see Brian L. Strom, *Basic Principles of Clinical Epidemiology Relevant to Pharmacoepidemiologic Studies*, in *Pharmacoepidemiology* 44, 50–55 (Brian L. Strom et al. eds., 6th ed. 2020), <https://doi.org/10.1002/9781119413431.ch3>.

26. Although epidemiologists refer generally to “health outcomes,” some disease is usually the health outcome of interest in the types of court cases in which epidemiologic research plays an important role. See *supra* discussion, note 1.

Figure 1. Design of a cohort study.



Cohort Studies

In cohort studies, researchers define a study population without regard to the participants' disease status. The cohort may be defined in the present and followed forward into the future (prospectively), in which case they are known as prospective, longitudinal, or follow-up studies. Alternatively, a cohort study may be constructed retrospectively as of some time period in the past and followed from that historical time toward the present. In cohort studies, there is usually a group that is unexposed and another that is exposed, or several groups with different levels of exposure (including, in appropriate cases, a group with no exposure). Exposure may also be measured continuously (e.g., blood lead levels). In a prospective study, individuals are enrolled and followed over time, during which exposure and subsequent health outcomes are measured, while in a retrospective study, the researcher will obtain data on past exposure from available records, samples, questionnaires, or other evidence.²⁷ So, for a study aiming to determine the association between an exposure and the occurrence of a disease, a researcher would compare the proportion of unexposed individuals who have the disease with the proportion of exposed individuals who have the

27. Sometimes in retrospective cohort studies, the researcher gathers historical data about exposure and disease outcome of a cohort. See Harold A. Kahn, *An Introduction to Epidemiologic Methods* 39–41 (1983); Mark A. Klebanoff & Jonathan M. Snowden, *Historical (Retrospective) Cohort Studies and Other Epidemiologic Study Designs in Perinatal Research*, 219 *Am. J. Obstetrics & Gynecology* 447 (2018), <https://doi.org/10.1016/j.ajog.2018.08.044>. Irving Selikoff, in his seminal study of asbestotic disease in insulation workers, included several hundred workers who had died before he began the study. Selikoff was able to obtain information about exposure from union records and information about disease from hospital and autopsy records. Irving J. Selikoff et al., *The Occurrence of Asbestosis Among Insulation Workers in the United States*, 132 *Ann. N.Y. Acad. Sci.* 139, 143 (1965), <https://doi.org/10.1111/j.1749-6632.1965.tb41097.x>.

disease.²⁸ If the exposure causes the disease, the researcher would expect a greater proportion of the exposed individuals to develop the disease than the unexposed individuals.²⁹ For a health outcome that is measured continuously, such as intelligence quotient (IQ) or blood pressure, the researcher could compare the average of these outcomes among the exposed and unexposed. If exposure is measured as a continuous rather than a dichotomous variable—for example, if all study subjects were exposed to some radiation from a nuclear accident but the degree of exposure was lower for those farther away from the accident—the researcher could examine the association between the extent of exposure and the health outcome.

One advantage of the cohort study design is that the temporal relationship between exposure and disease can often be established more readily than in other study designs. By tracking people who initially have not been diagnosed with the disease, the researcher can determine the time of disease onset (or diagnosis) and its relation to exposure. This temporal relationship is critical to the question of causation, because exposure must precede disease onset for exposure to have caused the disease.

As an example, in 1950 a cohort study was begun to determine whether uranium miners exposed to radon were at increased risk of death due to lung cancer as compared with nonminers. The study group (also referred to as the exposed cohort) consisted of 3,400 underground miners. The control group or unexposed cohort (which need not be the same size as the exposed cohort) comprised nonminers from the same geographic area. Members of the exposed cohort were examined every three years, and the degree of this cohort's exposure to radon was measured from samples taken in the mines. Ongoing testing for radioactivity and periodic medical monitoring of lungs permitted the researchers to examine whether disease was linked to prior work exposure to radiation and allowed them to discern the relationship between exposure to radiation and disease. Exposure to radiation was associated with the development of lung cancer in uranium miners.³⁰

The cohort design often is used in occupational studies such as the one just discussed. But because the design is not experimental, the investigator has no control over what other exposures a worker in the study may have had. Hence, an increased risk of disease among the exposed group may be caused by agents other than the exposure of interest. A researcher planning a study must attempt to

28. See Table 1, *infra*, in the section titled “Relative Risk.”

29. Researchers often examine the rate of disease or death in the exposed and control groups. The rate of disease or death entails consideration of the number developing disease within a specified period. All smokers and nonsmokers will, if followed for 100 years, die. Smokers will die at a greater rate than nonsmokers in the earlier years.

30. This example is based on a study description in Abraham M. Lilienfeld & David E. Lilienfeld, *Foundations of Epidemiology* 237–39 (2d ed. 1980). The original study is Joseph K. Wagoner et al., *Radiation as the Cause of Lung Cancer Among Uranium Miners*, 273 NEJM 181 (1965), <https://doi.org/10.1056/NEJM196507222730402>.

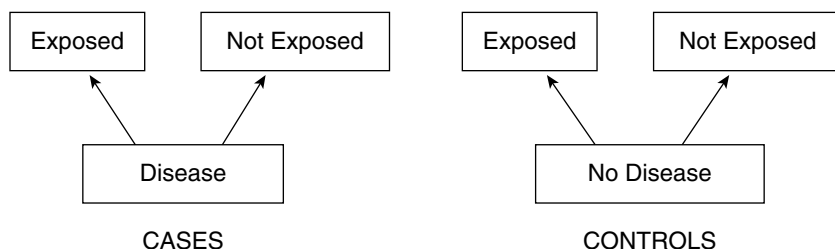
identify factors (called confounders) other than the exposure that may be related to exposure and may be responsible for the increased risk of disease. We discuss this problem (which is not limited to cohort studies) below in the “Confounding Bias” section. If data are gathered about potential confounders, the researcher generally uses statistical methods³¹ to estimate an association that is independent of these factors. Evaluating whether the association is causal involves additional analysis, as discussed below in the “General Causation” section.

Case-Control Studies

In case-control studies, the researcher begins with a group of individuals who have a disease (cases) and then selects a similar group of individuals who do not have the disease (controls). Controls should come from the same source population as the cases. The researcher then compares the extent of exposure in the two groups. For example, researchers might use employment records or questionnaire responses to determine whether and to what extent each study subject was exposed to the agent of interest, or researchers might measure biomarkers in the cases and controls to infer their exposure history. If a certain exposure is associated with the disease, a higher odds of exposure would be observed among the cases than among the controls (see Figure 2). Because researchers employing this research design often (though not always) gather historical information about exposure to an agent in the case and control groups, case-control studies have sometimes, but less precisely, been referred to as “retrospective studies.”³²

For example, in the late 1960s, doctors in Boston were confronted with an unusual number of young female patients with vaginal adenocarcinoma. Those

Figure 2. Design of a case-control study.



31. See generally Daniel L. Rubinfeld & David Card, *Reference Guide on Multiple Regression and Advanced Statistical Models*, in this manual; Kaye & Stern, *supra* note 16, “Statistical Models” (discussing difficulties of multiple regression).

32. As described, the design of a case-control study is inherently retrospective rather than prospective. But other types of epidemiologic studies, including some cohort studies, may also be retrospective. See Strom, *supra* note 25 and accompanying text.

patients became the cases in a case-control study (because they had the disease in question). Controls who did not have the disease were selected based on their being born in the same hospitals and at the same time as the cases. The cases and controls were compared for exposure to agents that might be responsible, and researchers found maternal ingestion of diethylstilbestrol (DES), a drug prescribed to prevent miscarriage and premature delivery, in all but one of the cases but none of the controls.³³

An advantage of the case-control study is that it usually can be completed in less time and with less expense than a cohort study. Case-control studies are also particularly useful in the study of rare diseases, because if a cohort study were conducted, a large group of individuals would have to be studied in order to observe the development of a sufficient number of cases for statistical analysis.³⁴ A number of potential limitations of case-control studies are discussed in the section below titled “Biases.”

Cross-Sectional Studies

A third type of observational study is a cross-sectional study. In this type of study, for each study participant the presence of both the exposure of interest and the health outcome of interest is assessed at a single point in time. Thus, cross-sectional studies reveal the prevalence (i.e., the presence at that particular time) of both exposure and disease but do not provide the disease incidence (i.e., the development of disease over time).³⁵ A researcher interested in the association between a health outcome (e.g., IQ) and an exposure that is relatively consistent over time (e.g., blood lead levels from living in contaminated housing) might use a cross-sectional study design. However, because a cross-sectional study determines both exposure and disease in an individual at the same point in time, it may be impossible to establish the temporal relation between exposure and disease—that is, whether the exposure preceded the disease, which is necessary for drawing any causal inference. Indeed, cross-sectional studies may be

33. See Arthur L. Herbst et al., *Adenocarcinoma of the Vagina: Association of Maternal Stilbestrol Therapy with Tumor Appearance in Young Women*, 284 NEJM 878 (1971), <https://doi.org/10.1056/NEJM197104222841604>.

34. For example, to detect a statistically significant doubling of disease caused by exposure to an agent where the incidence of disease is 1 in 100 in the unexposed population would require sample sizes of 3,100 for the exposed and nonexposed groups for a cohort study, but only 177 for the case and control groups in a case-control study (see section titled “Sampling Error” below for a discussion of statistical significance). Harold A. Kahn & Christopher T. Sempos, *Statistical Methods in Epidemiology* 66 (1989). See also Kaye & Stern, *supra* note 16, “What Is the Standard Error?” “What Is the Confidence Interval?” and “Tests or Interval Estimates?”

35. See *In re Deepwater Horizon Belo Cases*, No. 3:19CV963-MCR-HTC, 2022 WL 17721595, at *10 (N.D. Fla. Dec. 15, 2022) (criticizing failure of expert to identify a study on which she relied as a cross-sectional study).

affected by reverse causation, in which the health outcome under study actually causes a change in exposure.³⁶ Though more limited than cohort or case-control studies, cross-sectional studies can provide valuable leads to further directions for research, particularly when reverse causation can be ruled out based on other knowledge.³⁷

Ecological Studies

Up to now, we have discussed studies in which data on both exposure and health outcome are obtained for each individual included in the study, although individual data are sometimes augmented by group-level information, such as data from census tracts. Other studies, called ecological studies, collect only aggregate data on groups as a whole. In ecological studies, rather than gathering information about individuals, researchers obtain and compare overall rates of disease or death (or other summary measure for continuous health outcomes) for different exposure groups.³⁸

One type of ecological study compares exposure groups across geographic locations. For example, epidemiologists might compare disease rates among areas with more or less air pollution.³⁹ When making such comparisons, epidemiologists may overlay data on demographic factors obtained from sources such as the American Community Survey. Epidemiologists may look at geospatial or temporal trends (which may identify so-called “clusters” of disease) to form hypotheses and research questions.⁴⁰

Ecological studies may be useful to identify associations for further exploration, but they can rarely provide causal answers. We illustrate the difficulty of

36. For example, a cross-sectional study might find that disinfectant use is associated with greater odds of asthma, but could not distinguish whether disinfectants cause people to have asthma or whether having asthma causes people to use disinfectants.

37. For more information and references about cross-sectional studies, see David D. Celen-tano & Moyses Szklo, *Gordis Epidemiology* 154–57 (6th ed. 2018).

38. Studies may be conducted in which all members of a group or community are treated as exposed to an agent of interest (e.g., a contaminated water system) and disease status is determined individually. These studies should be distinguished from ecological studies.

39. See Jonah Lipsitt et al., *Spatial Analysis of COVID-19 and Traffic-Related Air Pollution in Los Angeles*, 153 *Env't Int'l* 106531 (2021), <https://doi.org/10.1016/j.envint.2021.106531>.

40. For example, the observed emergence of a cluster of adverse events associated with the use of heparin, a longtime and widely prescribed anticoagulant, led to suspicions that some specific lot of heparin was responsible. The observed pattern led the Centers for Disease Control and Prevention to conduct a case-control study, which concluded that contaminated heparin manufactured by Baxter was responsible for the outbreak of adverse events. See David B. Blossom et al., *Outbreak of Adverse Event Reactions Associated with Contaminated Heparin*, 359 *NEJM* 2674 (2008), <https://doi.org/10.1056/NEJMoa0806450>; *In re Heparin Prods. Liab. Litig.*, 803 F. Supp. 2d 712 (N.D. Ohio 2011).

using ecological studies to determine causality with the following hypothetical ecological study.

Suppose that researchers, interested in determining whether a high dietary-fat intake is associated with breast cancer, compared different countries in terms of their average fat intakes and their average rates of breast cancer. If countries with high average fat intake also tend to have high rates of breast cancer, the finding would suggest an association between dietary fat and breast cancer. However, such a finding would be far from conclusive, because it lacks particularized information about an individual's exposure and disease status (i.e., whether an individual with high fat intake is more likely to have breast cancer).⁴¹ In addition to the lack of information about an individual's intake of fat, the researcher does not know about the individual's exposures to other agents (or other factors, such as a woman's age when she first gave birth) that may also be responsible for the increased risk of breast cancer. This lack of information about each individual's exposure to an agent and disease status can lead to an erroneous inference about the relationship between fat intake and breast cancer, a problem known as the ecological fallacy. The fallacy is assuming that, on average, the individuals in the study who have suffered from breast cancer consumed more dietary fat than those who have not suffered from the disease. This assumption may not be true. Nevertheless, the study is useful in that it identifies an area for further research: the fat intake of individuals who have breast cancer as compared with the fat intake of those who do not. Researchers who identify a difference in disease or death in an ecological study may follow up with a study of individuals.⁴²

In another type of ecological study, epidemiologists may compare disease rates over time and focus on disease rates before and after a point in time when some event of interest took place. For example, after the once widely prescribed drug Bendectin was removed from the market, the rate of limb-reduction birth defects did not change, which suggested that Bendectin did not cause birth

41. For a discussion of the data on this question and what they might mean, see Kaye & Stern, *supra* note 16.

42. Some courts have admitted expert testimony that relied on the results of ecological studies. In *Cook v. Rockwell Int'l Corp.*, 580 F. Supp. 2d 1071, 1095–96 (D. Colo. 2006), the plaintiffs' expert conducted an ecological study in which he compared the incidence of two cancers among those living in a specified area adjacent to the Rocky Flats Nuclear Weapons Plant with other areas more distant. (The likely explanation for relying on this type of study is the time and expense of a study that gathered information about each individual in the affected area.) The court recognized that ecological studies are less probative than studies in which data are based on individuals but nevertheless held that this limitation went to the weight of the study. Plaintiffs' expert was permitted to testify to causation, relying on the ecological study he performed. See also *Mead v. Sec'y of Health & Hum. Servs.*, 2010 WL 892248, at *38 (Fed. Cl. Mar. 12, 2010) ("Because these studies examine trends in the aggregate rather than examining disease onset in exposed individuals, ecological studies are afforded less evidentiary weight in the medical community than controlled studies are.").

defects.⁴³ By contrast, researchers' discovery that the timing of a dramatic increase in limb-reduction birth defects followed widespread use of the drug thalidomide supported the conclusion that thalidomide caused those defects.⁴⁴

However, other than with such powerful agents as thalidomide, which increased the incidence of limb-reduction birth defects by several orders of magnitude, these secular-trend studies (also known as timeline studies) are less reliable and less able to detect modest causal effects than are the observational studies described above.⁴⁵ Other factors that affect the measurement or existence of the disease, such as improved diagnostic techniques and changes in lifestyle or age demographics, may change over time. If those factors can be identified and measured, it may be possible to control for them with statistical methods. Of course, unknown factors cannot be controlled for in these or any other kinds of epidemiologic studies.

Genetic and Molecular Epidemiologic Studies

Epidemiology, like other fields of biological and medical science, has in recent decades been touched powerfully by advances in scientists' ability to identify and study the molecular constituents of human cells, tissues, and organs. Perhaps the best-known of these advances is the development of genomics: the study of the human genome, the sequence of DNA bases arranged in chromosomes that form

43. See *Wilson v. Merrell Dow Pharms.*, 893 F.2d 1149 (10th Cir. 1990) (affirming judgment entered for defendant on a jury verdict). In *Wilson*, the defendant introduced evidence showing total sales of Bendectin and the incidence of limb-reduction birth defects during the 1970–1984 period. In 1983, Bendectin was removed from the market, but the rate of limb-reduction birth defects did not change. The Tenth Circuit held that the district court correctly determined that the timeline data were admissible and the defendant's expert witnesses could rely on those data in rendering their opinions. *Id.* at 1152–53. See also *Siharath v. Sandoz Pharms. Corp.*, 131 F. Supp. 2d 1347 (N.D. Ga. 2001). In *Siharath*, the court took note of the absence of secular trend data to support plaintiff's claim. After describing why several observational studies produced, at best, inconclusive results about any association between use of the lactation-suppressing drug Parlodel and the very rare complication of postpartum stroke, the court observed: "No evidence has been offered of an increase in postpartum strokes after the drug [Parlodel] was approved for suppression of lactation; no evidence has been offered of a decrease in postpartum strokes after the approval for suppression of lactation was withdrawn." *Id.* at 1358.

44. Michael D. Green, *Bendectin and Birth Defects* 70–71 (1996) (describing how thalidomide's teratogenicity was discovered after Dr. Widukind Lenz found a dramatic increase in the incidence of limb-reduction defects in Germany beginning in 1960).

45. See *Graddy v. Sec'y of Health & Hum. Servs.*, No. 08–416V, 2017 WL 11286536, at **18–22 (Fed. Cl. Aug. 31, 2017) (finding that the expert witness's ecological study on childhood autism prevalence and the introduction or increased doses of certain vaccines did not allow for a reliable evidentiary inference of causation regarding MMR II, varicella, and hepatitis A vaccines and autism).

the basis of heredity.⁴⁶ But genomics is just one of several fields of large-scale research on categories of biological molecules and features, such as epigenomics,⁴⁷ proteomics,⁴⁸ and others. Rapid technological and computational improvements have increasingly facilitated “omics” studies that can be performed on a larger scale, at a lower cost, and in a shorter time.⁴⁹

Variations in genes and other biochemical constituents across the population may be associated with the incidence of disease or the ability to metabolize an agent, and therefore can be subject to epidemiologic study. Genetic epidemiology assesses the genetic components that, perhaps in combination with environmental or other risk factors, contribute to the development or progression of disease in human populations.⁵⁰ More generally, molecular epidemiology “joins an understanding of disease at a molecular level with population-based study designs and approaches.”⁵¹ The number of epidemiologic studies that include molecular and genetic factors is quickly increasing.⁵²

Genetic and molecular epidemiologic studies add to epidemiology the ability to consider types of data that were previously unavailable. Where researchers previously could collect data only on the health outcome of interest, the exposure of interest, and relatively easily observed other potentially relevant traits

46. See Int’l Hum. Genome Sequencing Consortium, *Initial Sequencing and Analysis of the Human Genome*, 409 *Nature* 860 (2001), <https://doi.org/10.1038/35057062>. Genomics is contrasted to genetics by its scale: genetics is the study of the inheritance or effect of one gene or a relatively small number of genes; genomics is the study of the entire human genome or large portions of it.

47. Epigenetic factors are biochemical features or constituents that affect the extent to which genes are “turned on,” that is, expressed through synthesis of the proteins the genes encode. They have been described as an “extra layer of instructions that influences gene activity.” Gail P. A. Kauwell, *Epigenetics: What It Is and How It Can Affect Dietetics Practice*, 108 *J. Am. Dietetic Ass’n* 1056, 1056 (2008), <https://doi.org/10.1016/j.jada.2008.03.003>. Epigenomics has been defined as “[t]he study of epigenetic marks . . . on a genome-wide scale.” Pauline A. Callinan & Andrew P. Feinberg, *The Emerging Science of Epigenomics*, 15 *Hum. Molecular Genetics* R95, R96 (2006), <https://doi.org/10.1093/hmg/ddl095>.

48. Proteomics is the study of the proteome, the “complete set of proteins made by an organism.” Nat’l Cancer Inst., *Proteomics*, Dictionary of Cancer Terms, <https://perma.cc/74G5-ZWNB>; *Proteome*, *id.*, <https://perma.cc/ZQ79-HX5R>.

49. See, e.g., Matthew K. Sigurdson et al., *Redundant Meta-Analyses Are Common in Genetic Epidemiology*, 127 *J. Clinical Epidemiology* 40, 46 (2020), <https://doi.org/10.1016/j.jclinepi.2020.05.035> (“The advent of genomewide agnostic approaches and massive new biobanks and consortia has created a new paradigm for human genome epidemiology.”); Sara Goodwin et al., *Coming of Age: Ten Years of Next-Generation Sequencing Technologies*, 17 *Nature Rev. Genetics* 333, 333–37 (2016), <https://doi.org/10.1038/nrg.2016.49> (describing gene sequencing technologies and stating that cost of sequencing a single genome had been reduced to about \$1,000).

50. This description is closely paraphrased from John S. Witte & Duncan C. Thomas, *Genetic Epidemiology*, in Timothy L. Lash et al., *Modern Epidemiology* 963, 963 (4th ed. 2021), <https://doi.org/10.1093/aje/kwj102>.

51. Claire H. Pernar et al., *Molecular Epidemiology*, in Lash et al., *supra* note 50 at 935.

52. See, e.g., Muin J. Khoury et al., *Genetic Epidemiology with a Capital E, Ten Years After*, 35 *Genetic Epidemiology* 845 (2011), <https://doi.org/10.1002/gepi.20634>.

such as age or socioeconomic status, researchers may now also be able to collect information about the study subjects' genotype or relevant cellular or molecular components. But adding genetic or molecular information about the study subjects does not somehow turn epidemiology from a population-based science into the study of individuals, any more than including information about the study subjects' exposures and health outcomes did. Neither do genetic and molecular epidemiology constitute new types of epidemiologic study designs. Rather, they apply existing modes of epidemiologic study to genetic or molecular information. Thus, genetic and molecular epidemiologic studies employ various combinations of the study designs, such as cohort studies and case-control studies, described in the preceding subsection. Depending on the type of genetic or molecular data under study and the details of the study design, genetic and molecular epidemiologic studies are subject to the same sources of error that must be considered in designing and interpreting the results of any epidemiologic study.⁵³

As with other epidemiologic studies, genetic and molecular epidemiology studies may be used to support or undermine claims of disease causation in litigation. Although commentators have long forecast that the output of genetic and molecular epidemiology would revolutionize causal proof, as of this writing few judicial opinions have addressed these types of studies, and it is far from clear that a revolution is in the offing. Nevertheless, it is increasingly likely that expert witnesses will rely on such studies in forming their opinions presented in courts.

In particular, genetic epidemiology may provide additional evidence of whether there is a causal connection between an environmental exposure and a disease at issue in litigation—either in a general sense or in the plaintiff's specific instance. For example, a study found that maternal exposure to organophosphate insecticides was associated with shorter gestational duration only among infants who had a genetic variant that resulted in a slower elimination of these insecticides.⁵⁴ Alternatively, genetic epidemiology may reveal associations between genetic variations and a plaintiff's disease, raising the issue of whether or not a genetic variation may be a competing cause of the disease. This requires assessment of whether the gene-disease association is causal in a general sense, whether it acts independently of the exposure, and whether it is a competing cause in the plaintiff's specific instance. The extreme, though not typical, example would be

53. Pernar et al., *supra* note 51, at 936 (emphasizing that although molecular epidemiology involves new “omics” technologies, “molecular epidemiology can only contribute valid and reproducible findings if epidemiologists critically integrate core methodological concepts of epidemiology”); *id.* at 947–52 (discussing usefulness of various epidemiologic study designs in molecular epidemiology research).

54. Kim G. Harley et al., *Association of Organophosphate Pesticide Exposure and Paraoxonase with Birth Outcome in Mexican-American Women*, PLoS One 6(8):e23923 (2011), <https://doi.org/10.1371/journal.pone.0023923>.

a health outcome or disease entirely determined by genetics,⁵⁵ as is the case with sickle cell anemia.⁵⁶

Molecular epidemiology may also become relevant to causation disputes in ways that go beyond genetics. In molecular epidemiology studies, as one researcher described them, “risk factors, outcomes, confounders or effect modifiers are measured with biomarkers.”⁵⁷ Parties may use these types of studies to support or undermine claims that a plaintiff was exposed to an alleged toxicant, that an exposure caused a plaintiff’s disease or otherwise affected the plaintiff’s biochemistry or physiology, or that the plaintiff suffers from a particular condition or disease. In particular, parties may attempt to rely on research that seeks to identify “signature” molecular markers of disease etiology, with the validity, sensitivity, and specificity of the purported signatures likely to be in dispute.⁵⁸ For example, in several cases involving claims that exposure to benzene caused a plaintiff’s leukemia, courts have considered the significance of certain chromosome aberrations as possible biomarkers of exposure or effect.⁵⁹

55. *E.g.*, *Bowen v. E.I. Du Pont de Nemours & Co.*, No. CIV.A. 97C-06-194 CH, 2005 WL 1952859 (Del. Super. Ct. June 23, 2005), *aff’d*, 906 A.2d 787 (Del. 2006) (discussing how a newly discovered test for a newly discovered genetic variant established that a genetic mutation rather than a toxic exposure was “the cause of” plaintiff’s condition).

56. The sickle cell trait is a well-known example of a genetically determined condition. *See* Muin J. Khoury & Janice S. Dorman, *Genetic Disease*, in *Molecular Epidemiology* 365, 370 (Paul A. Schulte & Frederica P. Perera eds., 1993).

57. Paolo Boffetta, *Biomarkers in Cancer Epidemiology: An Integrative Approach*, 31 *Carcinogenesis* 121, 125 (2010), <https://doi.org/10.1093/carcin/bgp269>. Confounders are discussed in the section titled “Biases” below and effect modifiers are discussed in the section titled “Effect Modification” below.

58. Validity, in this context, refers to the degree to which the measurement of the “signature” marker measures what it purports to measure, as well as to whether the selected marker is generalizable from the study that derived it to the case in which it is being used. Sensitivity refers to the ability of an “etiologic signature” to identify true causal cases, i.e., to avoid false negatives. Specificity refers to the ability of an etiologic signature to identify false causal cases, i.e., to avoid false positives. The absence of a highly sensitive etiologic marker would be stronger evidence against causation than would be the absence of a less sensitive marker. The presence of a highly specific etiologic marker would be stronger evidence for causation than would be the presence of a less specific marker. *See* Steve C. Gold, *When Certainty Dissolves into Probability: A Legal Vision of Toxic Causation for the Post-Genomic Era*, 70 *Wash. & Lee L. Rev.* 237, 265–76 (2013) (discussing these concepts); *see also infra* Glossary of Terms.

59. *E.g.*, *Henricksen v. ConocoPhillips Co.*, 605 F. Supp. 2d 1142, 1149–50 (E.D. Wash. 2009) (noting absence in plaintiff of chromosomal aberrations found more frequently in toxicant-induced acute myelogenous leukemia than in idiopathic cases); *Hendrian v. Safety-Kleen Sys., Inc.*, No. 08-14371, 2014 WL 1464462, at *7 (E.D. Mich. Apr. 15, 2014) (describing plaintiff’s expert’s testimony that presence of chromosomal aberrations indicated prior exposure to benzene); *Harris v. KEM Corp.*, No. 85 CIV. 2127 (WK), 1989 WL 200446, at *3–*5 (S.D.N.Y. Dec. 2, 1989) (denying defendant’s motion for summary judgment where plaintiff’s expert testified that plaintiff’s leukemia had chromosomal aberrations consistent with benzene exposure).

Epidemiologic and Toxicologic Studies

In addition to observational epidemiology, toxicology models based on animal studies (in vivo) may be used to determine toxicity in humans.⁶⁰ Animal studies have a number of advantages. They can be conducted as true experiments by assigning exposure at random and carefully controlling and measuring exposure. Animal studies can avoid the other problems that human epidemiology studies confront such as confounding⁶¹ by, for example, genetics, social factors, or nutrition; participant refusal, non-compliance, or loss to follow-up; and certain ethical issues including those involving certain subpopulations, such as pregnant women and children. Test animals can be sacrificed and their tissues examined, which may improve the accuracy of disease assessment.⁶² Animal studies often provide useful information about pathological mechanisms and play a complementary role to epidemiology by assisting researchers in framing and assessing the plausibility of hypotheses.

Animal studies have two significant disadvantages, however. First, animal study results must be extrapolated to another species—human beings—and differences in absorption, metabolism, and other factors may result in interspecies variation in responses. For example, one powerful human teratogen, thalidomide, does not cause birth defects in most rodent species.⁶³ Similarly, some known teratogens in animals are not believed to be human teratogens.⁶⁴ The second difficulty with inferring causation in humans based on animal studies is that such studies often use doses that are substantially higher than concentrations to which humans are exposed. This may require the estimation of dose–response relationships for lower doses that have not been tested in order to establish a dose at which no effect is expected. For many agents, such no–effect thresholds may

60. For an in-depth discussion of toxicology, see David L. Eaton et al., *Reference Guide on Toxicology*, in this manual.

61. See section titled “Effect Modification” below.

62. Ethical considerations against deliberately exposing humans to agents thought harmful, see *supra* note 18, have in the past not been thought to prohibit animal experimentation. Contemporary concern for the ethical treatment of animals used in research is reflected in a number of statutory and regulatory provisions. See, e.g., 15 U.S.C. § 2603(h)(1) (2022) (amended Toxic Substances Control Act requires Environmental Protection Agency to “reduce and replace, to the extent practicable, scientifically justified, and consistent with the policies of [the Act], the use of vertebrate animals in the testing of chemical substances and mixtures”); Nat’l Insts. Health, *Animal Care & Use in the Intramural Research Program*, NIH Policy Manual § 3040-2 (Mar. 18, 2022), available at <https://perma.cc/9LW3-QJBQ>; *id.* Appendix 1 (listing applicable statutes, regulations, standards, and policies).

63. Phillip Knightley et al., *Suffer the Children: The Story of Thalidomide* 271–72 (1979).

64. In general, it is often difficult to confirm that an agent known to be toxic in animals is safe for human beings. See Ian C.T. Nisbet & Nathan J. Karch, *Chemical Hazards to Human Reproduction* 98–106 (1983); Int’l Agency for Research on Cancer, *Interpretation of Negative Epidemiological Evidence for Carcinogenicity* (Nicholas J. Wald & Richard Doll eds., 1985) [hereinafter IARC (Wald & Doll)].

not be known or may not exist.⁶⁵ Thus, inference conducted solely on the basis of animal studies is fraught with considerable uncertainty.⁶⁶

Toxicologists also use *in vitro* methods, in which human or animal tissue or cells are grown in laboratories and are exposed to certain substances. While useful, the problem with this approach is also extrapolation—whether one can generalize the findings from the artificial setting of tissues in laboratories to whole human beings.⁶⁷

Often toxicologic studies are the only or best available evidence of toxicity.⁶⁸ Epidemiologic studies are difficult, time-consuming, expensive, and

65. See *infra* text accompanying footnotes 235–37.

66. See *Soldo v. Sandoz Pharms. Corp.*, 244 F. Supp. 2d 434, 466 (W.D. Pa. 2003) (quoting the first edition of this reference guide); see also *Gen. Elec. Co. v. Joiner*, 522 U.S. 136, 143–45 (1997) (holding that the district court did not abuse its discretion in excluding expert testimony on causation based on expert’s failure to explain how animal studies supported expert’s opinion that agent caused disease in humans).

67. For a further discussion of these issues, see Eaton et al., *supra* note 60, “Extrapolation from Animal (In Vivo) and Cell (In Vitro) Research to Humans.” A number of courts have grappled with the role of animal studies in proving causation in a toxic substance case. One line of cases takes a very dim view of their probative value. For example, in *Johnson v. Arkema, Inc.*, 685 F.3d 452 (5th Cir. 2012), the appellate court concluded that the district court did not abuse its discretion in discounting an animal study because “studies of the effects of chemicals on animals must be carefully qualified in order to have explanatory potential for human beings.” *Id.* at 463 (quoting *Allen v. Pennsylvania Eng’g Corp.*, 102 F.3d 194, 197 (5th Cir. 1996)). See also *Becnel v. BP Expl. & Prod., Inc.*, No. CV 17-1758-SDD-EWD, 2021 WL 4444723, at *2 (M.D. La. Sept. 28, 2021) (noting that relying on animal studies, in the absence of epidemiologic studies, is . . . “of very limited usefulness”) (quoting *Brock v. Merrell Dow Pharms., Inc.*, 874 F.2d 307, 313 (5th Cir. 1989)); *In re Mirena IUD Prod. Liab. Litig.*, 169 F. Supp. 3d 396, 445 (S.D.N.Y. 2016) (holding that the expert’s reliance on animal studies, without a sound basis for extrapolating these studies to humans, is inadmissible). Other courts have been more amenable to the use of animal toxicology in proving causation. See *Metabolife Int’l, Inc. v. Wornick*, 264 F.3d 832, 842 (9th Cir. 2001) (holding that the lower court erred in *per se* dismissing animal studies, which must be examined to determine whether they are appropriate as a basis for causation determination); see also *In re Paoli R.R. Yard PCB Litig.*, 916 F.2d 829, 853–54 (3d Cir. 1990) (questioning the basis for the lower court’s exclusion of animal studies and remanding for further development of the record); *Drake v. Allergan, Inc.*, 2014 WL 12718976, at *1 (D. Vt. Oct. 23, 2014) (“The Court is mindful that animal studies present certain risks but these risks are not sufficient to exclude them categorically, especially where there is other evidence of causation as is the case here.”). The Third Circuit in a subsequent opinion in *Paoli* observed:

[I]n order for animal studies to be admissible to prove causation in humans, there must be good grounds to extrapolate from animals to humans, just as the methodology of the studies must constitute good grounds to reach conclusions about the animals themselves. Thus, the requirement of reliability, or “good grounds,” extends to each step in an expert’s analysis all the way through the step that connects the work of the expert to the particular case.

In re Paoli R.R. Yard PCB Litig., 35 F.3d 717, 743 (3d Cir. 1994); see also *Cavallo v. Star Enter.*, 892 F. Supp. 756, 761–63 (E.D. Va. 1995) (courts must examine each of the steps that lead to an expert’s opinion), *aff’d in part and rev’d in part*, 100 F.3d 1150 (4th Cir. 1996).

68. The International Association for Research on Cancer (IARC), a well-regarded international public health agency, evaluates the human carcinogenicity of various agents. In doing so,

sometimes—when it is difficult to accurately measure exposure, or when the exposure or the disease is extremely rare—virtually impossible to perform.⁶⁹ Consequently, epidemiologic studies do not exist for a large array of environmental agents. Where both toxicologic and epidemiologic studies are available, no universal rules exist for how to reconcile them.⁷⁰ Researchers often employ

IARC obtains, evaluates, and synthesizes all of the relevant evidence, including animal studies as well as any human studies. IARC then publishes a monograph containing that evidence, IARC's analysis, and a categorical assessment of the likelihood an agent is carcinogenic. In a preamble to each monograph, IARC explains what each of the categorical assessments means. IARC may classify a substance as “probably carcinogenic to humans” solely on the basis of the strength of animal studies. Int'l Agency for Research on Cancer, *Human Papillomaviruses*, 90 Monographs on Evaluation of Carcinogenic Risks to Humans 9–10 (2007), <https://perma.cc/J52S-ELWH> [hereinafter *IARC Papillomaviruses*]. When IARC monographs are available, courts generally recognize them as authoritative. See *Hardeman v. Monsanto Co.*, 997 F.3d 941, 967 (9th Cir. 2021) (affirming lower court's admission of IARC categorization of glyphosate as “a probable carcinogen”), *cert. denied*, 142 S. Ct. 2834 (2022). But IARC has only conducted evaluations of a fraction of potentially carcinogenic agents, and many suspected toxic agents cause effects other than cancer. See IARC, Revised Preamble for the IARC Monographs (2021), available at <https://perma.cc/XXY4-7DKF>.

69. Thus, in a series of cases involving Parlodel, a lactation suppressant for mothers of newborns, efforts to conduct an epidemiologic study of its effect on causing strokes were stymied by the infrequency of such strokes in women of child-bearing age. See, e.g., *Brasher v. Sandoz Pharms. Corp.*, 160 F. Supp. 2d 1291, 1297 (N.D. Ala. 2001); see also *In re Tylenol (Acetaminophen) Mktg., Sales Pracs. & Prods. Liab. Litig.*, No. 2:12-CV-07263, 2016 WL 3997046, at *7 (E.D. Pa. July 16, 2016) (in series of cases involving liver damage from use of acetaminophen (Tylenol) at or above the suggested dosage, efforts to conduct an epidemiological study of Tylenol's effect on causing acute liver failure (ALF) were stymied by the rarity of drug-induced ALF). In other cases, a plaintiff's exposure to an overdose of a drug may be unique or nearly so. See *Zuchowicz v. United States*, 140 F.3d 381 (2d Cir. 1998).

70. See IARC (Wald & Doll), *supra* note 64 (identifying several substances and comparing animal toxicology evidence with epidemiologic evidence). One explanation for the conflicting judicial treatment of toxicological studies of animals, see *supra* note 67, may be that when there is a substantial body of epidemiologic evidence that addresses the causal issue, courts find that animal toxicology has much less probative value. That was the case, for example, in several Bendectin cases. E.g., *Turpin v. Merrell Dow Pharms., Inc.*, 959 F.2d 1349, 1359 (6th Cir. 1992); *Brock v. Merrell Dow Pharms., Inc.*, 874 F.2d 307, 313; see also *In re Paoli R.R. Yard PCB Litig.*, No. 86-2229, 1992 U.S. Dist. LEXIS 16287, at *16 (E.D. Pa. 1992) (excluding evidence of or based on animal studies linking PCB exposure to various health effects). Where epidemiologic evidence is not available, animal toxicology may be thought to play a more prominent role in resolving a causal dispute. See Michael D. Green, *Expert Witnesses and Sufficiency of Evidence in Toxic Substances Litigation: The Legacy of Agent Orange and Bendectin Litigation*, 86 Nw. U. L. Rev. 643, 680–82 (1992) (arguing that plaintiffs should be required to prove causation by a preponderance of the available evidence). For another explanation of the Bendectin cases as well as other toxic tort case clusters, see Gerald W. Boston, *A Mass-Exposure Model of Toxic Causation: The Content of Scientific Proof and the Regulatory Experience*, 18 Colum. J. Env't L. 181 (1993) (arguing that epidemiologic evidence should be required in mass-exposure cases but not in isolated-exposure cases). See also IARC *Papillomaviruses*, *supra* note 68; Eaton et al., *supra* note 60, “Toxicology and Epidemiology.” The U.S. Supreme Court, in *General Electric Co. v. Joiner*, 522 U.S. 136, 144–45 (1997), suggested that there is no categorical rule for toxicologic studies, observing, “[W]hether animal studies can ever be a proper foundation for an expert's opinion [is] not the issue. . . . The [animal] studies were so dissimilar to

an approach that considers the overall weight of evidence, that is, all of the relevant scientific evidence that addresses the question of interest.⁷¹ This methodology entails making a judgment about causation after carefully assessing the results, validity, and consistency of each epidemiologic and toxicologic study.⁷²

Interpreting the Results of an Epidemiologic Study

Epidemiologists are ultimately interested in whether a causal link exists between an agent and a disease. However, the first question an epidemiologist generally addresses is whether an association is observed between exposure to the agent and disease. An association between exposure to an agent and disease is observed when disease occurs more frequently (or less frequently) when exposure exists than when exposure is absent.⁷³ Although a causal relation is one possible explanation for an observed association between an exposure and a disease, an association does not necessarily mean that there is a cause–effect relation. Interpreting the meaning of an observed association is discussed below.⁷⁴

the facts presented in this litigation that it was not an abuse of discretion for the District Court to have rejected the experts’ reliance on them.”

71. The methodology relies on expert scientific judgment to evaluate and weight the available evidence in order to reach the best conclusion. See Douglas L. Weed, *Weight of Evidence: A Review of Concept and Methods*, 25 Risk Analysis 1545 (2005), <https://doi.org/10.1111/j.1539-6924.2005.00699.x>. A number of courts have endorsed weight of evidence as a reliable methodology. See, e.g., *In re Deepwater Horizon Belo Cases*, No. 3:19CV963-MCR-HTC, 2022 WL 17721595, at *19 (N.D. Fla. Dec. 15, 2022) (stating that “weight of the evidence approach to analyzing causation can be considered reliable, provided the expert considers all available evidence carefully and explains how the relative weight of the various pieces of evidence led to his conclusion”) (quoting *In re Abilify (Aripiprazole) Prods. Liab. Litig.*, 299 F. Supp. 3d 1291, 1311 (N.D. Fla. 2018)). As the court in *In re Zantac (Ranitidine) Prod. Liab. Litig.*, 644 F. Supp. 3d 1075, 1168 (S.D. Fla. 2022), cogently observed: “Due to the ‘substantial judgment’ required of an expert in following this approach, it is crucial that the expert describe each step in the process by which he gathered and assessed the relevant scientific evidence” (quoting *Abilify*, *supra*, 299 F. Supp. at 1311)). When weight of evidence methodology is employed by a testifying expert, a critical step is adequately taking into account evidence contrary to the expert’s opinion. See *Norris v. Baxter Healthcare Corp.*, 397 F.3d 878, 882 (10th Cir. 2005) (“We are simply holding that where there is a large body of contrary epidemiological evidence, it is necessary to at least address it with evidence that is based on medically reliable and scientifically valid methodology.”); *Zantac*, *supra*, at *129 (“a plaintiff’s expert must address epidemiological evidence that is inconsistent with his or her causation opinions”).

72. See sections titled “Statistical Power” and “General Causation” below.

73. A negative association implies that the agent has a protective or curative effect. Because the concern in toxic substances litigation is whether an agent caused disease, this reference guide focuses on positive associations.

74. See section titled “General Causation” below.

This section begins by describing the ways of expressing the existence and strength of an association between exposure and disease or health outcome. It then proceeds to address sources of error that may produce an incorrect or skewed association, including random chance (sampling error) and non-random or systematic error (bias). Systematic error may result, for example, from inaccuracies in ascertaining health and exposure, from the manner of selecting those studied and information about them, or from confounding bias. This section also describes the statistical methods epidemiologists use to evaluate the importance of, and to address, these potential sources of error with respect to whether an association is real.

The strength of an association between exposure and a health outcome can be stated in various ways. For categorical outcomes, such as the occurrence of a disease, measures often employed include relative risk, odds ratio, attributable risk, or standardized mortality (or morbidity) ratio.⁷⁵ For continuous outcomes, such as IQ, researchers use other measures to describe the relationship between exposure and outcome. For example, statistical methods can be used to estimate a regression coefficient (sometimes called a beta coefficient) that represents the mean change in an outcome for a unit change in the level of exposure.⁷⁶ Regardless of the type of health outcome studied or the particular measure of association used, each of these measurements of association examines the degree to which the risk of an adverse health outcome changes with exposure to an agent in a population of individuals.

75. These are the type of health outcomes, and thus the measures of association, most commonly at issue in litigation. See, e.g., *In re Roundup Prods. Liab. Litig.*, 390 F. Supp. 3d 1102, 1120 n.18 (N.D. Cal. 2018) (identifying, explaining, and comparing relative risk and odds ratio); *In re Lipitor (Atorvastatin Calcium) Mktg., Sales Pracs. & Prods. Liab. Litig.*, No. MDL214MN-02502RMG, 2016 WL 827067, at *3 (D.S.C. Feb. 29, 2016) (discussing expert's selection of method to express magnitude of association from among relative risk, odds ratio, and risk difference). The Glossary of Terms *infra* defines two additional measures of strength of association for categorical outcomes: hazard ratio (see *Cooper v. Takeda Pharms. Am., Inc.*, 191 Cal. Rptr. 3d 67, 73–74, 78–79 (Cal. App. 2015) (describing expert testimony that defined and gave examples of hazard ratios)) and risk difference. A risk difference is the difference between the proportion of disease in those exposed to the agent and the proportion of disease in those who were unexposed. Thus, in the example given in the section titled “Relative Risk” below, the proportion of disease in those exposed is 40/100, the proportion of disease in the unexposed is 20/100, and the risk difference is 20/100 (40/100 minus 20/100). The risk difference is related to the attributable proportion of risk, another measure of association that is addressed below in the section titled “Attributable Risk.”

76. To estimate the association between an exposure and a continuous outcome, researchers compute beta coefficients, which are derived from standard multiple regression statistical analysis. See *infra* Glossary of Terms (defining beta coefficient). For more information on beta coefficients, see Rubinfield & Card, *supra* note 31, “Appendix.” Court opinions discussing epidemiology have not, to date, addressed beta coefficients.

Relative Risk

A commonly used approach for expressing the association between an agent and disease in cohort studies is the relative risk (RR).⁷⁷ It is defined as the ratio of the probability of disease in a group of exposed individuals (R_e in Table 1 and the equation below) to the probability of disease in a group of unexposed individuals (R_u in Table 1 and the equation below). Consider the information a researcher would obtain in conducting a cohort study as depicted in Table 1.

Table 1. Cross Tabulation of Exposure by Disease Status

	No Disease	Disease	Totals	Risk of Disease
Not Exposed	<i>a</i>	<i>c</i>	<i>a + c</i>	$R_u = c/(a + c)$
Exposed	<i>b</i>	<i>d</i>	<i>b + d</i>	$R_e = d/(b + d)$

The risk of disease is defined as the number of cases of disease that develop divided by the number of persons in the cohort under study.⁷⁸ Thus, the risk expresses the probability that a member of the population will develop the disease.

$$RR = \frac{R_e}{R_u} = \frac{\frac{d}{b + d}}{\frac{c}{a + c}}$$

For example, a researcher studies 100 individuals who are exposed to an agent and 200 individuals who are not exposed. After one year, 40 of the exposed individuals are diagnosed as having a disease, and 20 of the unexposed individuals also are diagnosed as having the disease. The relative risk of contracting the disease is calculated as follows:

- The risk of disease in the exposed individuals is 40 cases per 100 persons (40/100), or 0.4.
- The risk of disease in the unexposed individuals is 20 cases per 200 persons (20/200), or 0.1.

77. A relative risk cannot be calculated for a case-control study. An odds ratio is used instead to express the magnitude of any association found in such studies. See section titled “Odds Ratio” below.

78. Epidemiologists also use the concept of prevalence, which measures the existence of disease in a population at a given point in time, regardless of when the disease developed. Prevalence is expressed as the proportion of the population with the disease at the chosen time. See Celentano & Szklo, *supra* note 37, at 51–55.

- The relative risk is calculated as the ratio of the risk in the exposed group (0.4) divided by the risk in the unexposed group (0.1), or 4.0.

A relative risk of 4.0 indicates that the risk of disease in the exposed group is four times as high as the risk of disease in the unexposed group.⁷⁹

Risk, as defined above (generally referred to as cumulative incidence), can only be estimated if all individuals are followed for an equal period of time. If the follow-up time period differs between individuals in a study, the incidence rate is used instead. Incidence rates account for the fact that individuals followed for a longer period of time have a greater opportunity to develop the disease under study than those followed for a shorter time. To compute an incidence rate, a researcher divides the number of observed cases by the cumulated person-time.⁸⁰

In general, the relative risk can be interpreted⁸¹ as follows:

- If the relative risk equals 1.0, the risk in the group of exposed individuals is the same as the risk in the group of unexposed individuals.⁸² There is no association between exposure to the agent and disease.
- If the relative risk is greater than 1.0, the risk in the group of exposed individuals is greater than the risk in the group of unexposed individuals. There is a positive association between exposure to the agent and the disease, which may reflect a harmful effect of the agent in increasing the extent of disease.
- If the relative risk is less than 1.0, the risk in exposed individuals is less than the risk in unexposed individuals. There is a negative (or inverse) association between exposure to the agent and the disease, which could

79. See Celentano & Szklo, *supra* note 37, at 242–45; *Magistrini v. One-Hour Martinizing Dry Cleaning*, 180 F. Supp. 2d 584, 591 (D.N.J. 2002) (explaining the relationship between relative risk and the increased incidence of disease in the exposed group).

80. Person-time weights study participants by the amount of time those participants were observed and at risk for the disease. For example, one participant observed for one year is one person-year; one participant observed for two years is two person-years, etc. Two participants observed for one year each is also two person-years. The incidence rate is computed as the number of new cases of disease divided by the total person-time observed. See Celentano & Szklo, *supra* note 37, at 46–49.

81. For the sake of simplicity, this description refers only to relative risk. As noted in the immediately preceding text, if a cohort study follows different participants for different periods of time, the researcher computes an incidence rate ratio to determine if an association is observed. If an epidemiologic study reports an incidence rate ratio rather than a relative risk, the same interpretations as described in the text for relative risk would apply.

82. See *Magistrini*, 180 F. Supp. 2d at 591.

reflect a protective or curative effect of the agent on risk of disease. For example, immunizations lower the risk of disease. Thus, immunizations are expected to be negatively associated with the diseases that they target.

Although relative risk is a straightforward concept, care must be taken in interpreting it. Whenever an association is uncovered, further analysis should be conducted to assess whether the association is causal or due to chance (random error) or bias (including confounding).⁸³ These sources of error may also mask or skew a true association, resulting in a study that erroneously finds no association or overstates or understates the true association.

Epidemiologists often use relative risk as a measure of association in cohort studies. But researchers performing case-control studies cannot calculate a relative risk.⁸⁴ Instead, epidemiologists conducting case-control studies compute a different measure of association: an odds ratio.

Odds Ratio

The odds ratio (OR) is similar to a relative risk in that it expresses in quantitative terms the association between exposure to an agent and a disease. It is a convenient way to estimate an association, most particularly in case-control studies, though it can also be calculated for a cohort study.⁸⁵ The odds ratio approximates the relative risk when the disease is rare.⁸⁶

In a case-control study, the odds ratio is the ratio of the odds⁸⁷ that a case (one with the disease) was exposed to the odds that a control (one without the

83. See sections titled “Biases” and “Sampling Error” below.

84. A case-control study begins by examining a group of persons who already have the disease and another group of persons who do not have the disease. See section titled “Case-Control Studies” above. Because the number of cases and the number of controls are predetermined, the researcher cannot calculate the rate at which exposed and unexposed individuals in the study develop the disease. Lacking a rate or incidence of disease for both those exposed and those not exposed to the agent, a researcher cannot calculate a relative risk.

85. In a cohort study, the odds ratio is the ratio of the odds of a disease occurring when exposed to a suspected agent to the odds of the disease occurring when not exposed. An odds ratio can also be calculated for a cross-sectional study.

86. See Marcello Pagano & Kimberlee Gauvreau, *Principles of Biostatistics* 123 (3d ed. 2022). If the disease is not rare, the odds ratio is still valid to determine whether an association exists, but interpretation of its magnitude is less intuitive. For further detail about the odds ratio and its calculation, see Kahn & Sempos, *supra* note 34, at 47–56.

87. The odds of exposure is the ratio of those who were exposed to those who were not. Thus, from Table 2 *infra*, we can calculate the odds of exposure among the cases as a/c , and the odds of exposure among the controls as b/d .

disease) was exposed. Consider a case-control study, with results as shown schematically in a 2×2 table (Table 2):

Table 2. Cross Tabulation of Cases and Controls by Exposure Status

	Cases (with disease)	Controls (no disease)
Exposed	<i>a</i>	<i>b</i>
Not exposed	<i>c</i>	<i>d</i>

In a case-control study, the odds ratio *OR* is the odds that a case was exposed divided by the odds that a control was exposed. Looking at Table 2, this ratio can be calculated as:

$$OR = \frac{a/c}{b/d} = \frac{ad}{bc}$$

Because we are multiplying two diagonal cells in the table and dividing by the product of the other two diagonal cells, the odds ratio is sometimes also called the cross-products ratio.

Consider the following hypothetical study: A researcher identifies 100 individuals with a disease who serve as cases and 100 people without the disease who serve as controls for her case-control study. Of the 100 cases (with disease), 40 were exposed to an agent under study and 60 were not. Among the control group, 20 people were exposed and 80 were not. The data can be presented in a 2×2 table (Table 3).

Table 3. Case-Control Study Outcome

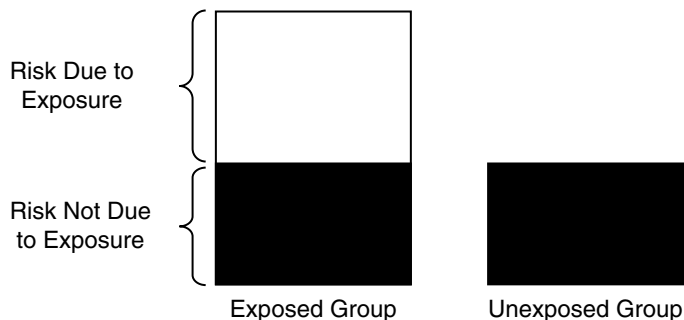
	Cases (with disease)	Controls (no disease)
Exposed	40	20
Not exposed	60	80

The calculation of the odds ratio would be $OR = (40/60)/(20/80) = 2.67$.

If the disease is relatively rare in the general population (about 5% or less), the odds ratio is a good approximation of the relative risk. Thus, in our example, there is almost a tripling in the risk of the disease in those exposed to the agent relative to those who were not exposed.⁸⁸

88. The odds ratio is usually more extreme (farther away from the null of 1.0) compared to the relative risk. In other words, if a relative risk is above 1.0, the odds ratio will tend to be larger; if a relative risk is below 1.0, the odds ratio will tend to be smaller. As the disease in question becomes more common, the difference between the odds ratio and the relative risk grows.

Figure 3. Risks in exposed and unexposed groups.



Attributable Risk

A frequently used measurement of risk is the attributable risk (AR), also called the attributable proportion of risk, the etiologic fraction, or the attributable risk percent. This measure represents the proportion of disease among exposed individuals that can be attributed to the exposure⁸⁹ and therefore, if the association is causal, the proportion of disease that could potentially be prevented if exposure to the agent were eliminated (see Figure 3).⁹⁰

The reason the odds ratio approximates the relative risk when the incidence of disease is small can be demonstrated by comparing the relative risk of a cohort study with its odds ratio. Referring to Table 1, the odds that the exposed group developed the disease is d/b ; similarly, the odds that the unexposed group developed the disease is c/a ; thus the odds ratio is $(d/b)/(c/a) = ad/bc$. Recall the relative risk would be $[d/(b+d)] / [c/(a+c)] = [d(a+c)] / [c(b+d)]$. Thus, when the incidence of disease is low, i.e., c and d are small in relation to a and b , the relative risk will approximate the odds ratio of ad/bc . See Celentano & Szklo, *supra* note 37, at 248–50.

If the disease is not rare, the relative risk may be estimated from the odds ratio. The formula for this estimate takes into account the prevalence of the disease among the unexposed (P_0):

$$RR = OR / (1 - P_0 + (P_0 \times OR)).$$

In our example, if the disease had a 15% prevalence among the unexposed, the corresponding estimated relative risk would be $2.67 / (1 - 0.15 + (0.15 \times 2.67)) = 2.14$. Note that, as described above, with a relatively high prevalence the numerical value of the odds ratio is farther from the null—in this case, greater than—the numerical value of the estimated relative risk. See Robert L. Grant, *Converting an Odds Ratio to a Range of Plausible Relative Risks for Better Communication of Research Findings*, 348 Brit. Med. J. f7540 (2014), <https://doi.org/10.1136/bmj.f7450>.

89. Kenneth J. Rothman et al., *Measures of Effect and Measures of Association*, in Lash, *supra* note 50, at 79, 93–97; see also *Landrigan v. Celotex Corp.*, 605 A.2d 1079, 1086 (N.J. 1992) (illustrating that a relative risk of 1.55 corresponds to an attributable risk of approximately 35%).

90. Risk is not zero for the control group (those not exposed) when there are other causal chains that cause the disease that do not require exposure to the agent. For example, some birth defects are the result of genetic sources, which do not require the presence of any environmental agent. Also, some degree of risk in the control group may be the result of background exposure to

To determine the proportion of a disease that is attributable to an exposure, a researcher would need to know the risk of the disease in the exposed group and the risk of disease in the unexposed group. The attributable risk is

$$AR = \frac{(\text{risk in the exposed}) - (\text{risk in the unexposed})}{\text{risk in the exposed}}$$

The attributable risk can be calculated using the example described in the section titled “Relative Risk.” Suppose a researcher studies 100 individuals who are exposed to a substance and 200 who are not exposed. After one year, 40 of the exposed individuals are diagnosed as having a disease, and 20 of the unexposed individuals are also diagnosed as having the disease.

- The risk of disease in the exposed group is 0.4 (40 persons with the disease out of 100 who are exposed).
- The risk of disease in the unexposed group is 0.1 (20 persons with the disease out of 200 who are not exposed).
- The proportion of disease in the exposed group that is attributable to the exposure is $(0.4 - 0.1)/0.4 = 0.75$ or 75%.

This means that 75% of the disease in the exposed group is attributable to the exposure. We should emphasize here that attributable does not necessarily mean “caused by.” Up to this point, we have only addressed associations. Inferring causation from an association is addressed in the “General Causation” section below.

Internal and External Validity

Internal validity is concerned with whether the result of a study—an association, for example in an epidemiologic context—is an accurate assessment of the true association in the source population from which study participants were recruited. Various biases such as selection bias, confounding bias, and information bias (all discussed in the “Biases” section below) can affect internal validity. For example, if researchers interviewing participants in a case-control study know the participant’s disease status, they might subconsciously probe harder to find exposure to the agent in participants with the disease than in those without

the agent being studied. For example, nonsmokers in a control group may have been exposed to passive cigarette smoke, which is responsible for some cases of lung cancer and other diseases. *See Ethyl Corp. v. EPA*, 541 F.2d 1, 25 (D.C. Cir. 1976) (describing the difficulty of finding people without exposure to ubiquitously distributed agents). There are some diseases that do not occur without exposure to an agent; these are known as signature diseases.

disease. That would produce an inflated estimate of the real association and would compromise the study's internal validity.⁹¹

External validity concerns whether the study results can be generalized to a different population. In the multidistrict litigation involving the drug phenylpropanolamine (PPA) used as an appetite suppressant, an epidemiologic study found an odds ratio of 16.58 for hemorrhagic stroke.⁹² The study was limited to individuals between the ages of 18 and 49 and, because of the paucity of men who had used PPA-containing appetite suppressants, the analysis was limited to women. In the litigation that followed, the court was faced with whether the study could be used to support causation for those under the age of 18, those over the age of 49, or men.⁹³ That issue called into question the matter of external validity: the extrapolation of the study results to a population that was not a part of the study.

Sources of Error in Epidemiologic Studies

Incorrect study results can occur in a variety of ways. A study may find a positive association (relative risk greater than 1.0) when there is no true association. Or a study may erroneously find that there is no association when in reality there is one. A study may also find an association when one truly exists, but the association found may be greater or less than the real association.

Two general categories of phenomena can cause erroneous associations: chance and bias. Before any inferences about causation are drawn from a study, the possible impact of these phenomena must be examined.⁹⁴

The findings of a study may be the result of chance (also called sampling error or random error). In designing a study, the size of the sample can be

91. Any problem with a study's internal validity would also affect the validity of any attempt to extrapolate the study's results to a different population. As described in the next paragraph, this would create a problem of external validity in addition to the impact on internal validity.

92. See W.N. Kernan et al., *Phenylpropanolamine and the Risk of Hemorrhagic Stroke*, 343 NEJM 1826 (Dec. 21, 2000), <https://doi.org/10.1056/NEJM200012213432501>.

93. See *In re Phenylpropanolamine (PPA) Prods. Liab. Litig.*, 289 F. Supp. 2d 1230, 1235 (W.D. Wash. 2003). There was additional evidence in the case that the court took into account; the description of the case in the text is a stylized one to concisely illustrate the idea of external validity.

94. See Philip Cole, *Causality in Epidemiology, Health Policy, and Law*, 27 Env't L. Rep. 10279, 10285 (1997); *DeLuca v. Merrell Dow Pharms., Inc.*, 911 F.2d 941, 955 (3d Cir. 1990) (recognizing and discussing random sampling error and then referring to other errors, such as systematic bias, that create as much or more error in the outcome of a study). See also *Daniels-Feasel v. Forest Pharms., Inc.*, No. 17 CV 4188-LTS-JLC, 2021 WL 4037820, at *2 (S.D.N.Y. Sept. 3, 2021) ("Where a positive association is observed, its validity is assessed by evaluating the role of possible alternative explanations, such as chance, bias, or confounding."). For a similar description of error in study procedure and random sampling, see Kaye & Stern, *supra* note 16, "What Inferences Can Be Drawn from the Data?"

increased to reduce (but not eliminate) the likelihood that a result may be due to chance. Statistical methods permit an assessment of the extent to which the results of a study may be due to chance, most often by computing a p -value and/or a confidence interval, which are closely related and are described in more detail in the “Sampling Error” section below.

Briefly, a calculated p -value is used to determine whether a study result is considered “statistically significant.” As described in more detail below, testing for statistical significance entails comparing the calculated p -value to a pre-selected threshold that enables an assessment of the role of random error in producing the study result. A result that is statistically significant is generally considered unlikely to be the result of chance, although any p -value threshold chosen is arbitrary.

A confidence interval provides information about the uncertainty (or precision) of an estimated measure of association (such as a relative risk or an odds ratio). A wide confidence interval suggests that chance is more likely to have an important impact on the estimate while a tight interval points to a likelihood of a smaller impact of chance and thus a more precise estimate.

We emphasize a point that those unfamiliar with statistical methodology frequently find confusing: That a study’s results are statistically significant says nothing about the magnitude of any association (i.e., the relative risk or odds ratio) or about the biological, clinical, or public health importance of the finding.⁹⁵ Significant, as used with the adjective statistically, does not mean “big” or “important.” A large study may find a statistically significant relationship that is quite modest in magnitude—for example, a relative risk of 1.05. Another study may find an association that is quite large—say, a relative risk of 10.0 or even higher—yet is not statistically significant because of the potential for random error due to a small sample size.⁹⁶ In short, *statistical significance is not about the size of the risk associated with an exposure.*⁹⁷

95. See Modern Scientific Evidence, *supra* note 3, § 5.36 at 435 (“Statisticians distinguish between ‘statistical’ and ‘practical’ significance. . . .”); Cole, *supra* note 94, at 10282. Suppose a study finds a statistically significant risk of 1.05. If that association applies to a rare, easily managed side effect of a medication prescribed to treat a common, debilitating illness, the finding might be more a reason for relief than for concern. But if 1.05 is the relative risk of a common, serious illness associated with ubiquitous exposure to an air pollutant, the finding might be cause for great concern despite the small magnitude of the relative risk. Alternatively, suppose a study finds a relative risk of 10.0 that is not statistically significant. If that relative risk applies to an invariably fatal condition associated with a small number of exposures to a new chemical that is poised to enter widespread use, the finding might suggest a need for caution or further investigation.

96. To find small effects that are statistically significant, larger sample sizes are required. When effects are larger, fewer subjects are required to produce statistically significant findings. For further explanation, see the section titled “Sampling Error” below.

97. Understandably, some courts have been confused about the relationship between statistical significance and the magnitude of the association. See, e.g., Hyman & Armstrong, P.S.C. v. Gunderson, 279 S.W.3d 93, 102 (Ky. 2008) (conflating the magnitude of the association with

Bias also can produce error in the outcome of a study. In epidemiology, as in science generally, bias refers to methodological issues that skew a study's results, with no connotation that a researcher has a subjective desire for, or interest in, a particular outcome. Bias encompasses any effect that tends to produce study results that differ from the true value in a systematic, rather than random, way. Thus, bias is also known as systematic error.

Epidemiologists attempt to minimize bias through their study designs, including data-collections protocols. Study designs are developed before researchers begin gathering data. However, some biases cannot be addressed at the design stage, and even the best-designed and conducted studies are likely to produce results that are biased to some extent. Consequently, after data collection is completed, statistical tools are often used to identify and correct for potential sources of bias. Epidemiologists may reanalyze a study's data to correct for a bias identified in a completed study or to validate the analytical methods used.⁹⁸ If correction (sometimes referred to as adjustment) for identified biases cannot be performed, epidemiologists can estimate whether the bias is likely to have inflated or diluted any association that may exist. Identification of uncorrected bias may enable an epidemiologist to make an assessment of the extent to which a study's conclusions are valid. Common biases and how they may produce invalid results are described in the "Biases" section.

Sampling Error⁹⁹

Before detailing the statistical methods used to assess sampling error (which we use synonymously with random error or chance), two concepts central to epidemiology and statistical analysis must be explained. Understanding these concepts should facilitate comprehension of the statistical methods. Epidemiologists often refer to true association (also called real association), which is the association that really exists between an agent and a disease in a given population and that might be found by a perfect (but nonexistent and impossible) study. The true association is a concept used in evaluating the results of a given study, even though its

whether it is statistically significant); *In re Pfizer Inc. Sec. Litig.*, 584 F. Supp. 2d 621, 634–35 (S.D.N.Y. 2008) (confusing the magnitude of the effect with whether the effect was statistically significant); *In re Joint E. & S. Dist. Asbestos Litig.*, 827 F. Supp. 1014, 1041 (S.D.N.Y. 1993) (concluding that any relative risk less than 1.50 is statistically insignificant), *rev'd on other grounds*, 52 F.3d 1124 (2d Cir. 1995). See generally Kaye & Stern, *supra* note 16, "Tests or Interval Estimates?"

98. E.g., Richard A. Kronmal et al., *The Intrauterine Device and Pelvic Inflammatory Disease: The Women's Health Study Reanalyzed*, 44 J. Clinical Epidemiology 109 (1991), [https://doi.org/10.1016/0895-4356\(91\)90259-C](https://doi.org/10.1016/0895-4356(91)90259-C) (a reanalysis of a study that found an association between the use of IUDs and pelvic inflammatory disease concluded that IUDs do not increase the risk of pelvic inflammatory disease).

99. For a bibliography on the role of statistical significance in legal proceedings, see Sanders, *supra* note 16, at 329 n.138.

value is unknown. By contrast, a study's outcome will produce an observed association, which is known.

Formal procedures for statistical testing begin with the null hypothesis, which posits that there is no true association (i.e., a relative risk of 1.0) between exposure to the agent and the disease under study. Data are gathered and analyzed to see whether they disprove the null hypothesis.¹⁰⁰ The data are subjected to statistical testing to assess the plausibility that any association found is a result of random error or whether, instead, the data support rejection of the null hypothesis.

The use of the null hypothesis for this testing should not be understood as the a priori belief of the investigator. When epidemiologists investigate an agent, it is usually because they hypothesize that the agent is a cause of some outcome. Nevertheless, in preparing their study designs, epidemiologists rely on the null hypothesis to test the plausibility that any association found in a study was the result of random error.¹⁰¹

False Positives and Statistical Significance

When a study results in a positive association (i.e., a relative risk greater than 1.0), epidemiologists try to determine whether that result represents a true association or is the result of random error. Random error can be illustrated by considering a fair coin (i.e., not modified to produce more heads than tails, or vice versa). On average, we would expect that coin tosses would yield half heads and half tails. But sometimes, a set of coin tosses might yield an unusual result—for example, six heads out of six tosses, an occurrence that, purely by chance, would result in less than 2% of a series of six tosses. In the world of epidemiology, sometimes the study findings, merely by chance, do not reflect the true relationships between an agent and outcome. Any single study—even a clinical trial—is in some ways analogous to a set of coin tosses, being subject to the play of chance. So, for example, even if the true relative risk (in the total population) were 1.0, an epidemiologic study might find an observed relative risk (in the study sample) greater than (or less than) 1.0 because of random error (i.e., chance).¹⁰² An erroneous conclusion that the null hypothesis is false (i.e., a conclusion that there is a different risk among those exposed to an agent relative to those that were

100. See, e.g., *Daubert v. Merrell Dow Pharms., Inc.*, 509 U.S. 579, 593 (1993) (scientific methodology involves generating and testing hypotheses).

101. See *DeLuca v. Merrell Dow Pharms., Inc.*, 911 F.2d 941, 945 (3d Cir. 1990); *United States v. Philip Morris USA, Inc.*, 449 F. Supp. 2d 1, 706 n.29 (D.D.C. 2006); Stephen E. Fienberg et al., *Understanding and Evaluating Statistical Evidence in Litigation*, 36 *Jurimetrics J.* 1, 21–24 (1995), [https://doi.org/10.1016/S0015-7368\(91\)73152-6](https://doi.org/10.1016/S0015-7368(91)73152-6).

102. See *Magistrini v. One-Hour Martinizing Dry Cleaning*, 180 F. Supp. 2d 584, 592 (D.N.J. 2002) (citing the second edition of this reference guide).

unexposed when no difference actually exists) owing to random error is called a false-positive error (also called a Type I error or alpha error).

Common sense leads one to believe that a large enough sample of individuals must be studied if the study is to identify a relationship that truly exists between exposure to an agent and a disease. Common sense also suggests that by enlarging the sample size (the size of the study group), researchers can form a more accurate conclusion and reduce the chance of random error in their results. Both statements are correct and can be illustrated by a test to determine if a coin is fair. A test in which a fair coin is tossed 1,000 times is more likely to produce close to 50% heads than a test in which the coin is tossed only 10 times. By contrast, it is far more likely that a test of a fair coin with 10 tosses will by chance come up with (for example) 80% heads than will a test with 1,000 tosses. With large numbers, the outcome of the test is less likely to be influenced by random error, and the researcher would have greater confidence in the inferences drawn from the data.¹⁰³

Given a set of epidemiologic data, one might want to ask the straightforward, obvious question: What is the probability that an observed association reflects a real association that exists in the population from which the study subjects were drawn? But traditional statistical methods cannot answer this question.¹⁰⁴ Instead, researchers compute “*p*-values” that address a related but very different question: Assuming there really is no association in the population—referred to as the null hypothesis—how probable is it that one would find an association in the study sample at least as large as the observed association?¹⁰⁵

103. This explanation of numerical stability was drawn from Brief Amicus Curiae of Prof. Alvan R. Feinstein in Support of Respondent, *Daubert v. Merrell Dow Pharms., Inc.*, No. 92-102, 1993 WL 13006284, at *12–*13 (U.S. Jan. 19, 1993). See also *Allen v. United States*, 588 F. Supp. 247, 417–18 (D. Utah 1984), *rev’d on other grounds*, 816 F.2d 1417 (10th Cir. 1987) (observing that although “[s]mall communities or groups of people are deemed ‘statistically unstable’” and “data from small populations must be handled with care[, it] does not mean that [the data] cannot provide substantial evidence in aid of our effort to describe and understand events”); Shi En Kim, *Heads or Tails*, Sci. Am., Jan. 2024, at 12 (explaining that if a coin were tossed a million times, an observed deviation of even one percent, such as 510,000 heads and 490,000 tails, would very likely result from a small but real bias in the coin).

104. For a discussion of Bayesian statistics, which under certain conditions can provide information about this probability, see *infra* text accompanying footnotes 120 and 178 and the entry “Bayesian analysis” in the Glossary of Terms.

105. This methodology, known as hypothesis testing, is one of the most counterintuitive techniques in statistics. See *Modern Scientific Evidence*, *supra* note 3, § 5.36 at 359 (“it is easy to mistake the *p*-value for the probability that there is no difference”).

The *p*-value for a given study does not provide a rate of error or even a probability of error for an epidemiologic study. In *Daubert*, 509 U.S. at 593, the Court stated that “the known or potential rate of error” should ordinarily be considered in assessing scientific reliability. Epidemiology, however, unlike some other methodologies—fingerprint identification, for example—does not permit

A p -value represents the probability that an association equal to or more extreme than (i.e., greater for a positive association or smaller for an inverse association) the observed association could be found *if no association were in fact present*, that is, if the null hypothesis were true.¹⁰⁶ Thus, a p -value of 0.1 means that there is a 10% chance that, with no association being actually present in the population from which the study sample was drawn (i.e., under the null hypothesis), values at least as large as an observed relative risk above 1.0 could have occurred by random chance.¹⁰⁷

The smaller the p -value, the less likely it is that the null hypothesis is true. Epidemiologists use a convention that the p -value must fall below some specified threshold, known as alpha or the significance level, for the results of the study to be considered statistically significant.¹⁰⁸ Thus, an outcome is statistically significant when the observed p -value for the study falls below the preselected significance level.

Significance testing allows researchers to minimize false positive results. The most common significance level used in science is 0.05. As explained above, a p -value of 0.05 means that the probability of observing an association at least as large as that found in the study, when in truth there is no association, is

an assessment of its accuracy by testing against a known reference standard. A p -value provides information only about the plausibility of random Type I error given the study result, but the true relationship between agent and outcome remains unknown. A p -value provides no information about the plausibility of random Type II error (false negatives), which is discussed in the sections titled “False Negatives” and “Statistical Power” below. Moreover, a p -value provides no information about whether other sources of error exist and, if so, their magnitude. See sections titled “Biases” and “Conceptual Error” below. In short, for epidemiology, there is no way to determine a rate of error—even for the random error aspect of studies. See *Kumho Tire Co. v. Carmichael*, 526 U.S. 137, 151 (1999) (recognizing that for different scientific and technical inquiries, different considerations will be appropriate for assessing reliability); *Cook v. Rockwell Int’l Corp.*, 580 F. Supp. 2d 1071, 1100 (D. Colo. 2006) (“Defendants have not argued or presented evidence that . . . there is, in fact, a method by which an overall ‘rate of error’ can be calculated for an epidemiologic study.”).

106. See also Kaye & Stern, *supra* note 16, “ p -values, Significance Levels, and Hypothesis Tests” (the p -value reflects the implausibility of the null hypothesis).

107. Technically, a p -value of 0.1 means that if in fact there is no association, if the same study were to be repeated, 10% of such studies would be expected to yield an association the same as, or greater than, the one found in the study due to random error. The interpretation is similar for observed relative risks below 1.0: a p -value of 0.10 would represent a 10% chance that values equal or smaller would be due to random error. Note that the description here and in the text describes a “one-tailed” significance test. For an explanation of the distinction between one-tailed and two-tailed tests, see *infra* note 113.

108. *Cook*, 580 F. Supp. 2d at 1100–01 (discussing p -values and their relationship with statistical significance); *Allen v. United States*, 588 F. Supp. 247, 416–17 (D. Utah 1984), *rev’d on other grounds*, 816 F.2d 1417 (10th Cir. 1987) (discussing statistical significance and selection of a level of alpha); see also Sanders, *supra* note 16, at 343–44 (explaining alpha, beta, and their relationship to sample size); *Developments in the Law—Confronting the New Challenges of Scientific Evidence*, 108 Harv. L. Rev. 1481, 1535–36, 1540–46 (1995) [hereinafter *Developments in the Law*].

equal to 5%.¹⁰⁹ When examining effect modification¹¹⁰ (such as when an association may differ by sex, social class, or genetics), epidemiologists usually use a higher p -value, e.g., a p -value of 0.10. This helps compensate for the fact that statistical power¹¹¹ to detect effect modification is lower relative to that of the power to detect overall associations because stratifying by an effect modifier effectively reduces the sample size.¹¹²

Figure 4 presents a graphical depiction to assist understanding of significance testing. Under the null hypothesis, we assume that, if we performed many similar studies identical to the one for which we wish to estimate a p -value and graphed the frequency of their outcomes, we would expect to observe a normal (bell-shaped) distribution, with most results clustered near the null value (remember, we assume the null hypothesis) and only a few extreme results in the “tails,” far from the hypothesized null value, as depicted by the small shaded

109. This means that if one examined a large number of situations in which the true relative risk equals 1—meaning there is no association between any of the agents studied and the outcome of interest—on average 1 result in 20 nevertheless would show an association that is statistically significant at a .05 level. When researchers examine many possible associations that might exist in their data—known as data dredging—we should expect that even if there are no true causal relationships, those researchers will find statistically significant associations in 1 of every 20 associations examined. See Rachel Nowak, *Problems in Clinical Trials Go Far Beyond Misconduct*, 264 Sci. 1538, 1539 (1994), <https://doi.org/10.1126/science.8202708>. For an accessible discussion of this problem, see Andrew Gelman & Eric Loken, *The Statistical Crisis in Science*, 107 Am. Scientist (2014), <https://doi.org/10.1511/2014.111.460>.

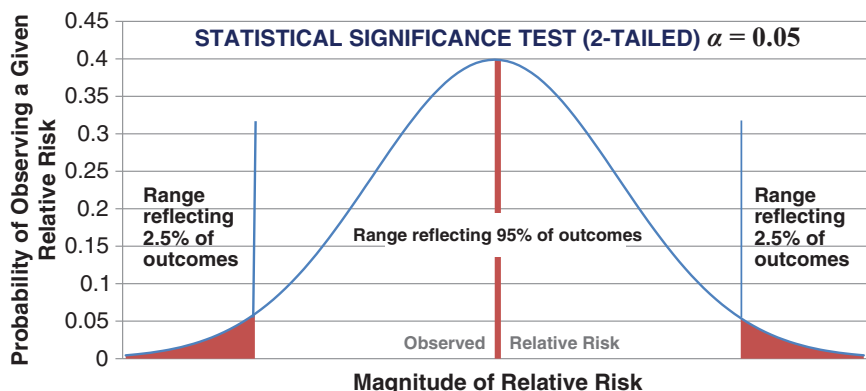
In certain contexts, the sheer number of tests—and therefore the number of associations that would appear by chance—is so large that researchers apply much more stringent conventions for statistical significance. For example, researchers testing hundreds of thousands of genetic variations for association with disease typically accept as statistically significant only results that would appear by chance no more than once in every 20 million trials. See Laura Buzdugan et al., *Assessing Statistical Significance in Multivariable Genome Wide Association Analysis*, 32 Bioinformatics 1990, 1991 (2016) (“the multiple testing issue is resolved by applying a stringent significance threshold—most commonly 5×10^{-8} ”). Some researchers favor even greater stringency. Sara L. Pulit et al., *Resetting the Bar: Statistical Significance in Whole-Genome Sequencing-Based Association Studies of Global Populations*, 41 Genetic Epidemiology 145 (2016), <https://doi.org/10.1002/gepi.22032> (arguing for significance levels ten to fifty times more stringent). Even in this context, the “cut points” chosen for statistical significance “are somewhat arbitrary and do not reflect the potential clinical or biological importance of associations, nor do they account for the cost of false negatives.” John S. Witte & Duncan C. Thomas, *Genetic Epidemiology*, in Lash et al., *supra* note 50, at 963, 970. For discussion of false negatives, see section titled “False Negatives” below.

110. See the section titled “Effect Modification” below for further explanation.

111. See the section titled “Statistical Power” below.

112. As explained *infra* at text accompanying footnote 172, researchers study effect modification using stratification—separating the study participants into subgroups (strata). For example, 200 participants in a study might be stratified into two strata of 100 males and 100 females. Because the number of participants in each stratum is smaller than the size of the study as a whole, the study’s statistical power for each stratum is reduced. The effect is illustrated below in an example at the end of the section titled “Effect Modification.”

Figure 4. Hypothetical probability distribution under the null hypothesis for an estimated relative risk of 4.0.



ranges in Figure 4. A significance test is ordinarily two-tailed, as depicted in Figure 4.¹¹³

There is some controversy among epidemiologists and biostatisticians about the appropriate role of significance testing.¹¹⁴ To the strictest significance testers,

113. Two-tailed p -values assume that the direction of the association, either greater or less than the null of 1.0, is not known, and thus distribute the probabilities on both sides of the null. For example, if a study estimated a relative risk of 4.0, a two-tailed p -value would include the probability that, if the null hypothesis is true, the study would find a relative risk either equal to or larger than 4.0 (the observed value) or equal to or less than 0.25 (the inverse of the observed value, i.e., $1/4.0$). A one-tailed test, which is sometimes employed, considers only probabilities on the same side of the null as the study result. For a given alpha, a one-tailed test effectively doubles the probability of finding a significant effect in the hypothesized direction and is therefore controversial among epidemiologists and uncommon in epidemiologic studies. See *In re Phenylpropanolamine (PPA) Prods. Liab. Litig.*, 289 F. Supp. 2d 1230, 1241 (W.D. Wash. 2003) (accepting the propriety of a one-tailed test for statistical significance in a toxic-substance case); *United States v. Philip Morris USA, Inc.*, 449 F. Supp. 2d 1, 701 (D.D.C. 2006) (explaining the basis for EPA's decision to use a one-tailed test in assessing whether secondhand smoke was a carcinogen). But see *Good v. Fluor Daniel Corp.*, 222 F. Supp. 2d 1236, 1243 (E.D. Wash. 2002) (concluding that, in the circumstances presented, a one-tailed test on which the expert relied was not valid). For more on the difference between one-tailed and two-tailed tests, see Kaye & Stern, *supra* note 16.

114. In 2019, *The American Statistician*, a journal of the American Statistical Society, published an editorial that contained numerous "don'ts" in using statistical significance that many had previously violated. R. Wasserstein et al., *Moving to a World Beyond $p < 0.05$* , 73 Am. Statistician 1 (Supp. 1, 2019). Some courts, contrary to the editorial's prescriptions, have used statistically significant results as a threshold for acceptability of a study. The leading case advocating statistically significant studies is *Brock v. Merrell Dow Pharms., Inc.*, 874 F.2d 307, 312 (5th Cir.), *amended*, 884 F.2d 167

any study whose p -value is greater than the level chosen for statistical significance should be rejected as inadequate to disprove the null hypothesis. However, the consensus among most epidemiologists is that conclusions about potential

(5th Cir. 1989). Overturning a jury verdict for the plaintiff in a Bendectin case, the court observed that no statistically significant study had been published that found an increased relative risk for birth defects in children whose mothers had taken Bendectin. A number of courts have followed the *Brock* decision or have indicated strong support for significance testing as a screening device. See, e.g., *In re Zolof* (Sertraline Hydrochloride) Prod. Liab. Litig., 26 F. Supp. 3d 449, 456, 465 (E.D. Pa. 2014) (ruling that an expert witness's reliance on non-statistically significant data was "not derived from principles and techniques of uncontroverted validity" and was a "departure from use of well-established epidemiological methods"); *Wagoner v. Exxon Mobil Corp.*, 813 F. Supp. 2d 771, 800 (E.D. La. 2011) ("An opinion on general causation is inadmissible if it rests entirely on studies that do not show statistically significant results.").

By contrast, a number of courts appear more cautious about using significance testing as a necessary condition, instead recognizing that assessing the likelihood of random error is important in determining the probative value of a study. In *Allen*, 588 F. Supp. at 417, the court stated, "The cold statement that a given relationship is not 'statistically significant' cannot be read to mean there is no probability of a relationship." The Third Circuit described confidence intervals (i.e., the range of values that would be found in similar studies as a result of chance, with a specified level of confidence) and their use as an alternative to statistical significance in *DeLuca v. Merrell Dow Pharms., Inc.*, 911 F.2d 941, 948–49 (3d Cir. 1990). See also, e.g., *In re Lipitor* (Atorvastatin Calcium) Mktg., Sales Prac. & Prods. Liab. Litig., 892 F.3d 624, 642 (4th Cir. 2018) ("[W]e decline to establish a bright-line rule requiring experts to rely only on evidence that is statistically significant or else have their opinions excluded."); *Milward v. Acuity Specialty Prod. Grp., Inc.*, 639 F.3d 11, 24–25 (1st Cir. 2011) ("the district court read too much into the paucity of statistically significant epidemiological studies" since it would be "very difficult to perform an epidemiological study of the causes of [the disease] that would yield statistically significant results").

In *Daubert v. Merrell Dow Pharms., Inc.*, 509 U.S. 579 (1993), although the trial court had relied in part on the absence of statistically significant epidemiologic studies, the Supreme Court did not explicitly address the matter. One commentator concluded that "*Daubert* did not set a threshold level of statistical significance either for admissibility or for sufficiency of scientific evidence." *Developments in the Law*, *supra* note 108, at 1535–36, 1540–46. The Supreme Court in *General Electric Co.*, 522 U.S. at 145–47, adverted to the lack of statistical significance in one study relied on by an expert as a ground for ruling that the district court had not abused its discretion in excluding the expert's testimony. In a different context—securities fraud—the Supreme Court held that a lack of statistical significance in adverse event reports did not disqualify them from being "material and requiring disclosure." *Matrixx Initiatives, Inc. v. Siracusano*, 563 U.S. 27 (2011). In the course of its opinion, the Court explained:

A lack of statistically significant data does not mean that medical experts have no reliable basis for inferring a causal link between a drug and adverse events. As *Matrixx* itself concedes, medical experts rely on other evidence to establish an inference of causation. We note that courts frequently permit expert testimony on causation based on evidence other than statistical significance. We need not consider whether the expert testimony was properly admitted in those cases, and we do not attempt to define here what constitutes reliable evidence of causation. It suffices to note that, as these courts have recognized, "medical professionals and researchers do not limit the data they consider to the results of randomized clinical trials or to statistically significant evidence."

Matrixx, 563 U.S. at 40–41 (citations and footnote omitted).

associations should not rely on strict significance testing¹¹⁵ and that it is inappropriate either to reject all results that fail to meet the specified p -value or to accept all results that do meet this level.¹¹⁶ Epidemiologists have become increasingly sophisticated in addressing the issue of random error and examining the data from a study that may provide information about the relation between an agent and a disease, without the necessity of basing their conclusions solely on statistical significance.¹¹⁷ Meta-analysis, a method for pooling the results of multiple studies, sometimes can ameliorate concerns about random error present in a single study.¹¹⁸

Calculation of a confidence interval permits a more refined assessment of appropriate inferences about the association found in an epidemiologic study.¹¹⁹ A confidence interval represents a range of possible values calculated from the

115. In addition to criticism of significance testing, many lawyers, judges, and legal academics have incorrectly stated that using an alpha of 0.05 for statistical significance testing imposes a burden of proof on the plaintiff far higher than the civil burden of a preponderance of the evidence (i.e., greater than 50%). See Michael D. Green, *Science Is to Law as the Burden of Proof Is to Significance Testing*, 37 *Jurimetrics J.* 205, 221 n.67 (1997) (book review) (citing many examples).

To fully explain why this comparison is mistaken would require more space and detail than is feasible here. We sketch out a brief explanation. First, using an alpha of 0.50 would not be equivalent to saying that the probability that the association found is real is 50%, and the probability that it is a result of random error is 50%. Statistical methodology does not permit assessment of those probabilities. Second, significance testing only bears on whether the observed magnitude of association arose as a result of random chance, not on whether the null hypothesis is true. Third, using stringent significance testing to avoid false-positive error comes at a complementary cost of inducing false-negative error. Fourth, alpha does not address the likelihood that a plaintiff's disease was caused by exposure to the agent; the magnitude of the association bears on that question. See section titled "Specific Causation" below. See Green, *supra* note 70, at 686–89; see also David H. Kaye, *Apples and Oranges: Confidence Coefficients and the Burden of Persuasion*, 73 *Cornell L. Rev.* 54, 66 (1987); Kaye & Stern, *supra* note 16, "Evaluating Hypothesis Tests"; *Turpin v. Merrell Dow Pharms., Inc.*, 959 F.2d 1349, 1357 n.2 (6th Cir. 1992); *cf.* *DeLuca v. Merrell Dow Pharms., Inc.*, 911 F.2d 941, 959 n.24 (3d Cir. 1990) ("The relationship between confidence levels and the more likely than not standard of proof is a very complex one . . . and in the absence of more education than can be found in this record, we decline to comment further on it.").

116. For a hypercritical assessment of statistical significance testing by two economists that nevertheless identifies much inappropriate overreliance on it, see Stephen T. Ziliak & Deidre N. McCloskey, *The Cult of Statistical Significance* (2008).

117. See Sanders, *supra* note 16, at 342 (describing the improved handling and reporting of statistical analysis in studies of Bendectin after 1980).

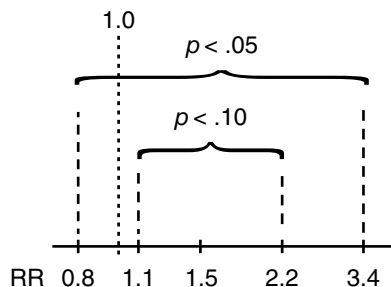
118. See section titled "Methods for Synthesizing or Combining the Results of Multiple Studies" below.

119. Kenneth Rothman, Professor of Public Health at Boston University and Adjunct Professor of Epidemiology at the Harvard School of Public Health, is one of the leaders in advocating the use of confidence intervals and rejecting strict significance testing. See *DeLuca v. Merrell Dow Pharms., Inc.*, 911 F.2d 941, 947 (3d Cir. 1990) (discussing Rothman's views on the appropriate level of alpha and the use of confidence intervals); *Turpin v. Merrell Dow Pharms., Inc.*, 959 F.2d 1349, 1353–54 n.1 (6th Cir. 1992) (discussing the relationship among confidence intervals, alpha, and statistical power). See also *Cook v. Rockwell Int'l Corp.*, 580 F. Supp. 2d 1071, 1100–01 (D. Colo. 2006) (discussing confidence intervals, alpha, and significance testing). For a discussion of

results of a study. For example, a 95% confidence interval signifies that 95% of the confidence intervals generated by multiple identical studies would include the true value. The width of the interval reflects random error. The narrower the confidence interval, the more statistically stable the results of the study.

The advantage of a confidence interval is that it displays more information than significance testing. Statistically significant does not convey the magnitude of the association found in the study or reveal how statistically stable that association is. A confidence interval shows boundaries of the relative risk (or other measure of association) based on selected levels of alpha or statistical significance. Just as the p -value does not provide the probability that the risk estimate found in a study is correct, the confidence interval does not provide the range within which the true risk is likely to lie. In other words, it is a misconception to interpret a 95% confidence interval as representing an interval within which the true value has a 95% probability of being found. Though some types of statistical analyses (referred to as Bayesian statistics) can generate intervals with this interpretation (referred to as credible intervals), most published research does not.¹²⁰ An example of two confidence intervals that might be calculated for a given relative risk is displayed in Figure 5.

Figure 5. 90% and 95% confidence intervals for a hypothetical study with a relative risk of 1.5.



The confidence intervals shown in Figure 5 are for a hypothetical study that found a relative risk of 1.5, with boundaries of 0.8 to 3.4 when alpha is set to 0.05 (equivalently, a confidence level of 0.95), and with boundaries of 1.1 to 2.2 when alpha is set to 0.10 (equivalently, a confidence level of 0.90). The confidence interval for alpha equal to 0.10 is narrower because it encompasses only

the use of confidence intervals in evaluating sampling error more generally than in the epidemiologic context, see Kaye & Stern, *supra* note 16, “What Is the Confidence Interval?”

120. See Jonathan A.C. Sterne & George Davey Smith, *Sifting the Evidence: What’s Wrong with Significance Tests*, 322 *BMJ* 226, 228 (2001) (comparing frequentist significance testing with Bayesian statistical approaches).

90% of the expected intervals. By contrast, the confidence interval for alpha equal to 0.05 includes the expected outcomes for 95% of the intervals. To generalize this point, the lower the alpha chosen (and therefore the more stringent the exclusion of positive results as possible random error), the wider the confidence interval. At a given alpha, the width of the confidence interval is determined by sample size. All other things being equal, the larger the sample size, the narrower the confidence boundaries (indicating greater numerical stability or precision). For a given risk estimate, a narrower confidence interval reflects a decreased likelihood that an association found in a particular study would occur by chance if the true association is 1.0.¹²¹

For the example in Figure 5, the boundaries of the confidence interval with alpha set at 0.05 encompass a relative risk of 1.0, and the result would be said not to be statistically significant at the 0.05 level. Alternatively, if the confidence boundaries are defined as an alpha equal to 0.10, then the confidence interval no longer includes a relative risk of 1.0, and the result could be described as statistically significant at the 0.10 level. Common practice in epidemiologic studies is to estimate 95% confidence intervals.

False Negatives

As Figure 5 illustrates, false positives can be reduced by adopting more stringent values for alpha, which increases the likelihood that a result will fail the selected test of statistical significance. Using an alpha of 0.05 will result in fewer false positives than using an alpha of 0.10. An alpha of 0.01 or 0.001 would produce even fewer false positives.¹²² The tradeoff for reducing false positives is an increase in false-negative errors (also called beta errors or Type II errors¹²³). This concept reflects the possibility that a study will be interpreted as negative (providing no grounds to reject the null hypothesis), when in fact there is a true

121. Where multiple epidemiologic studies are available, a technique known as meta-analysis may be used to combine the results of the studies (*see* section titled “Methods for Synthesizing or Combining the Results of Multiple Studies” below), thereby reducing the numerical instability of all the studies. *See generally* Diana B. Petitti, *Meta-Analysis, Decision Analysis, and Cost-Effectiveness Analysis: Methods for Quantitative Synthesis in Medicine* (2d ed. 2000).

122. As described above, *supra* note 109, in genome-wide association studies, which search the entire human genome to see if variations in genes are associated with variable disease risk, alpha is commonly set at 5×10^{-8} , allowing one false positive per 20 million tests. Some researchers, worried that such stringent significance testing causes too many real associations to be rejected as potential chance results, have proposed ameliorative approaches. *See* Shrahyashi Biswas et al., *A Framework for Pathway Knowledge Driven Prioritization in Genome-wide Association Studies*, 44 *Genetic Epidemiology* 841 (2020), <https://doi.org/10.1002/gepi.22345>.

123. *See* A Dictionary of Epidemiology 99 (Miquel Porta et al. eds., 6th ed. 2014).

association of a specified magnitude.¹²⁴ The probability of such an error, commonly called beta in statistical analyses, can be calculated for any study.

Statistical Power

When a study fails to find a statistically significant association, an important question is whether the result tends to exonerate the agent's toxicity or is essentially inconclusive with regard to toxicity.¹²⁵ The concept of statistical power can be helpful in evaluating whether a study's results, if not statistically significant, are exonerative or inconclusive.¹²⁶

The power of a study is the probability of finding a statistically significant association of a given magnitude (if it exists). The power of a study depends on several factors: the sample size; the level of alpha (or statistical significance) specified; the background incidence of disease; and the specified magnitude of the association (i.e., the size of the relative risk) that the researcher would like to detect.¹²⁷ Power curves can be constructed that show the likelihood of finding

124. See also *DeLuca v. Merrell Dow Pharms., Inc.*, 911 F.2d 941, 947 (3d Cir. 1990); *In re Testosterone Replacement Therapy Prods. Liab. Litig. Coordinated Pretrial Proc.*, No. 14 C 1748, 2017 WL 1833173, at *12 n.12 (N.D. Ill. May 8, 2017) (concluding expert's testimony comparing likelihood of Type I and Type II errors "will be helpful to the jury").

125. Even when a study or body of studies tends to exonerate an agent, that does not establish that the agent is absolutely safe. See *Cooley v. Lincoln Elec. Co.*, 693 F. Supp. 2d 767 (N.D. Ohio 2010). Epidemiology is not able to provide such evidence.

126. See Fienberg et al., *supra* note 101, at 22–23; Sander Greenland, *Null Misinterpretation in Statistical Testing and Its Impact on Health Risk Assessment*, 53 Preventative Med. 225 (2011), <https://doi.org/10.1016/j.jymed.2011.08.010> (explaining that statistical nonsignificance does not necessarily provide evidence for the null hypothesis and identifying such errors in reasoning). Thus, in *Smith v. Wyeth-Ayerst Labs. Co.*, 278 F. Supp. 2d 684, 693 (W.D.N.C. 2003) and *Cooley v. Lincoln Elec. Co.*, 693 F. Supp. 2d 767, 773 (N.D. Ohio 2010), the courts recognized that the power of a study was critical to assessing whether the failure of the study to find a statistically significant association was exonerative of the agent or inconclusive. See also *Procter & Gamble Pharms., Inc. v. Hoffmann-LaRoche Inc.*, No. 06 CIV. 0034 (PAC), 2006 WL 2588002, at *32 n.16 (S.D.N.Y. Sept. 6, 2006) (discussing power curves and quoting the second edition of this reference guide); *In re Phenylpropanolamine (PPA) Prods. Liab. Litig.*, 289 F. Supp. 2d 1230, 1243–44 (W.D. Wash. 2003) (explaining expert's testimony that "statistical reassurance as to lack of an effect would require an upper bound of a reasonable confidence interval close to the null value"); *Ruff v. Ensign-Bickford Indus., Inc.*, 168 F. Supp. 2d 1271, 1281 (D. Utah 2001) (explaining why study should be treated as inconclusive rather than exonerative based on small number of subjects in the study).

127. See Malcolm Gladwell, *How Safe Are Your Breasts?*, New Republic, Oct. 24, 1994, at 22, 26, <https://perma.cc/83AP-4AGM>; Kenneth J. Rothman & Timothy L. Lash, *Precision and Study Size*, in Lash et al., *supra* note 50, at 333, 354–55. For continuous health outcomes such as child IQ, the magnitude of the association is measured differently (see *supra* text accompanying footnotes 75–76 and *infra* entry "beta coefficient" in the Glossary of Terms), and instead of background incidence the statistical power would depend in part on the background variability of the outcome.

any given relative risk in light of these factors. Often, power curves are used in the design of a study to determine what size a study population should be.¹²⁸

The power of a study is the complement of beta ($1-\beta$). Thus, a study with a likelihood of 0.25 of failing to detect a true relative risk of 2.0¹²⁹ or greater has a power of 0.75. This means the study has a 75% chance of detecting a true relative risk of 2.0. If the power of a null study to find a relative risk of 2.0 or greater is low, it has substantially less probative value than a study with similar results but a higher power.¹³⁰

Choosing the sample size is the main way in which a researcher can affect the power of the study. Because studies with larger sample sizes are more powerful than studies with smaller sample sizes (everything else being equal), a null study with a large sample size is given more credence than a similar null study with a smaller sample size.¹³¹

Biases

Other than sampling error, another reason an epidemiologic study might find an invalid or inaccurate association is systematic error or bias, as distinguished from random error. Bias may arise in the design or conduct of a study, data collection, or data analysis. Some biases (such as confounding, described below) may also be inherent to the natural relation between the agent of interest and the disease under study.

The meaning of scientific bias differs from conventional (and legal) usage, in which bias refers to a partisan point of view.¹³² When scientists use the term *bias*, they refer to anything that results in a systematic (nonrandom) error in a study result and thereby compromises the study's validity. The statistical measures that address possible sampling error—significance tests, confidence intervals, and power—do not address systematic error (bias), because bias does not result from random chance in the selection of study subjects.¹³³

128. For examples of power curves, see Kenneth J. Rothman, *Modern Epidemiology* 80 (1986); Pagano & Gauvreau, *supra* note 86, at 239. For a different graphical presentation showing how a smaller study size results in a wider 95% confidence interval for a given level of absolute and relative risk, see Rothman & Lash, *supra* note 127, at 360.

129. We use a relative risk of 2.0 for illustrative purposes because of the legal significance courts have attributed to this magnitude of association. See section titled “Specific Causation” below.

130. See also Kaye & Stern, *supra* note 16; section titled “Effect Modification” below.

131. See *id.* for a more detailed discussion of statistical power.

132. See A Dictionary of Epidemiology, *supra* note 123, at 21; Edmond A. Murphy, *The Logic of Medicine* 239–62 (1976).

133. See Green, *supra* note 70, at 667–68; Vincent M. Brannigan et al., *Risk, Statistical Inference, and the Law of Evidence: The Use of Epidemiological Data in Toxic Tort Cases*, 12 *Risk Analysis* 343, 344–45 (1992), <https://doi.org/10.1111/j.1539-6924.1992.tb00686.x>. Contrary to the statement in

Most epidemiologic studies have some degree of bias that may affect the outcome. If major bias is present, it may invalidate the study results. Identifying biases, however, can be challenging and requires expertise in both epidemiologic methods as well as the subject matter at hand. In reviewing the validity of an epidemiologic study, the epidemiologist should identify potential biases and analyze the amount or kind of error that might have been induced by the bias. In some cases, the potential direction of the bias can be determined and, depending on the specific type of bias, it may exaggerate the real association, dilute it, completely mask it, or even reverse its direction.

Biases are generally classified in three categories: selection bias, information bias, and confounding bias. Selection bias can result from differences between individuals selected for a study and the larger population about which researchers want to make inferences based on the study results. Information bias can arise from errors in measuring exposure, disease, or other relevant variables. Confounding bias can occur if other causal factors influence the association between exposure to an agent and a disease.

Selection Bias

Selection bias refers to the error in an observed association that results from the method of selection of study participants, such as cases and controls (in a case-control study) or exposed and unexposed individuals (in a cohort study). Selection bias can occur in several ways. One common source of selection bias is the existence of non-random differences between individuals who are selected as study participants and those who are not.¹³⁴

The selection of an appropriate control group has been described as the Achilles' heel of a case-control study.¹³⁵ Controls should be drawn from the same

at least one early toxic tort case, *see, e.g., Brock v. Merrell Dow Pharms., Inc.*, 874 F.2d 307, 312 (5th Cir. 1989), a study's use of confidence intervals cannot "incorporate the possibility" of bias in the data. A statistically significant result with a narrow 95% confidence interval could be incorrect if bias exists. *See In re Zolof (Sertraline Hydrochloride) Prods. Liab. Litig.*, 858 F.3d 787, 793 (3d Cir. 2017) ("If a causal connection does not actually exist, significant findings can still occur due to, *inter alia*, inability to control for a confounding effect or detection bias.").

134. In *In re "Agent Orange" Prod. Liab. Litig.*, 597 F. Supp. 740, 783 (E.D.N.Y. 1985), *aff'd*, 818 F.2d 145 (2d Cir. 1987), the court expressed concern about selection bias. The exposed cohort consisted of young, healthy men who served in Vietnam. Comparing the mortality rate of the exposed cohort and that of a control group made up of civilians might have resulted in error that was a result of selection bias. Failing to account for health status as an independent variable tends to understate any association between exposure and disease in studies in which the exposed cohort is healthier. *See also In re Baycol Prods. Litig.*, 532 F. Supp. 2d 1029, 1043 (D. Minn. 2007) (upholding admissibility of testimony by expert witness who criticized study based on selection bias).

135. William B. Kannel & Thomas R. Dawber, *Coffee and Coronary Disease*, 289 NEJM 100 (1973) (editorial), <https://doi/full/10.1056/NEJM197410242911703>.

source population (a concept often referred to as the study base) that produced the cases. Doing so avoids the possibility that the controls will have different risk factors for the disease than the cases. Selecting control participants becomes problematic if the control participants are selected for reasons that are related to their exposure (or lack thereof) to the agent being studied. For example, a study of the effect of smoking on heart disease will suffer selection bias if subjects of the study are volunteers and the decision to volunteer is affected by both being a smoker (perhaps smokers are more likely to choose to participate in a study due to concerns about their health) and having a family history of heart disease (which increases the probability that they will suffer from the disease due to genetics). In this case, the association will be biased upward because smokers with higher probabilities of heart disease will be more likely to self-select for the study.

Hospital-based studies, which are relatively common among researchers located in medical centers, illustrate the problem. Suppose an association is found between coffee drinking and coronary heart disease in a study using hospital patients as controls. The problem is that the hospitalized control group may include individuals who had been advised against drinking coffee for medical reasons, such as to prevent the aggravation of a peptic ulcer. In other words, the controls may become eligible for the study because of their medical condition, which is in turn related to their exposure status—their likelihood of avoiding coffee. If this is true, the amount of coffee drinking in the control group would understate the extent of coffee drinking expected in people who do not have the disease and thus would bias upwardly (i.e., exaggerate) any odds ratio observed.¹³⁶ Bias in hospital studies may also understate the true odds ratio when the exposures at issue increase the likelihood of cases being hospitalized and also contribute to the controls' chances of hospitalization.

Just as cases and controls in case-control studies should be selected independently of their exposure status, the exposed and unexposed participants in cohort studies should be selected independently of their disease risk.¹³⁷ For example, if women with hysterectomies are overrepresented among exposed (smoking) women in a cohort study of the connection between smoking and cervical

136. Hershel Jick et al., *Coffee and Myocardial Infarction*, 289 NEJM 63 (1973), <https://doi.org/10.1056/NEJM197307122890203> (finding an association between coffee drinking and coronary heart disease using hospital patients for controls).

137. Selection bias can introduce confounding factors (see section titled “Effect Modification” below) when unexposed controls may differ from the exposed cohort because exposure is associated with other risk (or protective) factors. Investigators can attempt to measure and adjust for those differences, as explained in the section titled “Effect Modification” below. See also Martha J. Radford & JoAnne M. Foody, *How Do Observational Studies Expand the Evidence Base for Therapy?*, 286 JAMA 1228 (2001), <https://doi.org/10.1001/jama.286.10.1228> (discussing the use of propensity analysis to adjust for potential confounding and selection biases that may occur from nonrandomization in observational epidemiology).

cancer, this could understate the association between the exposure and the disease. This is so because women who have a hysterectomy cannot develop cervical cancer so the overrepresentation of women with hysterectomies would reduce the number of smoking women who develop cervical cancer in the study. If exposure to the agent is also related to participation in the study, selection bias could occur. For example, if women who are more health-conscious (and therefore are less likely to smoke) were also more likely to enroll in the study, a spurious association would be created between smoking and cervical cancer.

A further source of selection bias occurs when those selected to participate decline to participate or drop out before the study is completed. Many studies have shown that individuals who choose to participate or remain in studies over time differ substantially from those who do not. If a substantial portion of individuals declines to participate, the researcher should investigate whether those who declined are different from those who agreed. If data are available on non-participants, the researcher can compare relevant characteristics of those who participate with those who do not, to show the extent to which the two groups are comparable. Similarly, if a significant number of subjects drop out of a study before completion, the remaining subjects may not be representative of the original study populations. The researcher should examine whether that is the case.

The fact that a study may suffer from selection bias does not necessarily invalidate its results. A number of factors may suggest that a bias, if present, had only a limited effect. If the association is particularly strong, for example, bias is less likely to account for all of it. In addition, a consistent association across different control groups suggests that possible biases applicable to a particular control group are not invalidating. Similarly, a dose-response relationship (see section titled “Dose-Response Relationship” below) found among multiple groups exposed to different doses of the agent may provide additional evidence that biases applicable to the exposed group are not a major problem. However, the converse may not necessarily be true as the risk of disease associated with some agents may not consistently increase with exposure level, a concept referred to as nonlinearity.¹³⁸ Finally, it is important to note that if data are available about nonparticipants or those lost to follow-up, statistical methods can be applied to mitigate selection bias.

Information Bias

Information bias is a result of inaccurate information primarily about either the disease or the exposure status of the study participants.¹³⁹ In assessing whether the data may reflect inaccurate information, one must assess whether the data

138. See *infra* notes 237–41 and accompanying text.

139. Incorrect information about potential confounders can also bias the results of a study.

were collected from objective and reliable sources. Medical records, government documents, employment records, death certificates, and interviews are examples of data sources that are used by epidemiologists to measure either exposure or disease status. The accuracy of a particular source may affect the validity of a research finding. For example, using employment records to gather information about exposure to narcotics probably would lead to inaccurate results, because employees tend to keep such information private. If the researcher uses an unreliable source of data, the study may not be useful.

The kinds of quality-control procedures used may affect the accuracy of the data. For data collected by interview, quality-control procedures should probe the reliability of the individual and whether the information is verified by other sources whenever possible. For data collected and analyzed in the laboratory, quality-control procedures should assess the validity and reliability of the laboratory tests.

Exposure information bias

Though information bias may affect the results of a study of any study design, this category of bias (exposure information bias) is an important consideration for retrospective studies, such as case-control studies, where researchers depend on information from the past to determine exposure.¹⁴⁰ In some situations, the researcher may be required to interview the subjects to determine past exposures, thus relying on the subjects' memories. Research has shown that individuals with disease (cases) tend to recall past exposures more readily than individuals with no disease (controls);¹⁴¹ this creates a potential for a type of information bias called recall bias.

For example, consider a case-control study conducted to examine the cause of congenital malformations. The epidemiologist is interested in whether the malformations were caused by an infection during the mother's pregnancy. A group of mothers of infants with malformations (cases) and a group of mothers of infants with no malformation (controls) are interviewed regarding infections

140. Information bias can be a problem in cohort studies as well. When exposure is determined retrospectively, there can be a variety of impediments to obtaining accurate information. Similarly, when disease status is determined retrospectively, bias is a concern. The determination that asbestos is a cause of mesothelioma was hampered by inaccurate death certificates that identified lung cancer rather than mesothelioma, a rare form of cancer, as the cause of death. See I.J. Selikoff et al., *Mortality Experience of Insulation Workers in the United States and Canada*, 220 *Annals N.Y. Acad. Sci.* 91, 110–11 (1979), <https://doi.org/10.1111/j.1749-6632.1979.tb18711.x>; David E. Lilienfeld & Paul D. Gunderson, *The "Missing Cases" of Pleural Malignant Mesothelioma in Minnesota, 1979–81: Preliminary Report*, 101 *Pub. Health Rep.* 395, 397–98 (1986).

141. Steven S. Coughlin, *Recall Bias in Epidemiological Studies*, 43 *J. Clinical Epidemiology* 87 (1990).

during pregnancy. Mothers of children with malformations may recall an inconsequential fever or runny nose during pregnancy that readily would be forgotten by a mother who had a normal infant.¹⁴² Even if in reality the infection rate in mothers of children with malformations is no different from the rate in mothers of normal children, the result in this study would be an apparently higher rate of infection in the mothers of the children with the malformations solely on the basis of recall differences between the two groups.¹⁴³

The issue of recall bias can sometimes be addressed by finding a second source of data to validate the subject's response (e.g., blood-test results from pre-natal visits or medical records that document symptoms of infection).¹⁴⁴ Alternatively, the mothers' responses to questions about other exposures that were not

142. Recall bias among parents whose babies had birth defects was a prominent issue in Bendectin cases. See, e.g., *Brock*, 874 F.2d at 311–12 (discussing recall bias among women who bear children with birth defects); *Lynch v. Merrell-National Labs.*, 830 F.2d 1190, 1195 (1st Cir. 1987) (discussing a study's attempt to “wash out ‘maternal-recall bias’”).

143. In a case similar to this hypothetical, *Newman v. Motorola, Inc.*, 218 F. Supp. 2d 769, 778 (D. Md. 2002), the court considered a study of the effect of cell phone use on brain cancer and concluded that there was good reason to suspect that recall bias affected the results of the study, which found an association between cell phone use and cancers on the side of the head where the cell phone was used but no association between cell phone use and overall brain tumors. The court excluded plaintiff's proposed expert testimony on causation. *Id.* at 783.

144. Two researchers who used a case-control study to examine the association between congenital heart disease and the mother's use of drugs during pregnancy corroborated interview data with the mother's medical records. See Sally Zierler & Kenneth J. Rothman, *Congenital Heart Disease in Relation to Maternal Use of Bendectin and Other Drugs in Early Pregnancy*, 313 NEJM 347, 347–48 (1985). Courts have considered researchers' varied efforts to determine the extent to which recall bias affected study results. For example, in one case plaintiffs' experts “pointed to studies that sought to validate self-reports of pesticide exposure and that found similar recall accuracy between cases and controls.” *In re Roundup Prods. Liab. Litig.*, 390 F. Supp. 3d 1102, 1121 (N.D. Cal. 2018). In multidistrict litigation alleging that prolonged perineal use of talcum powder products caused ovarian cancer, an expert witness for plaintiffs acknowledged the possibility that recall bias affected the results of case-control studies, but opined that the possibility could be discounted because similar results were found in data collected both before and after widespread media coverage of the purported causal link. *In re Johnson & Johnson Talcum Powder Mktg., Sales Pracs., & Prods. Litig.*, 509 F. Supp. 3d 116, 167 (D.N.J. 2020). The court concluded that “Defendants and their experts may disagree with Plaintiffs' general causation experts' assessment of recall bias in the case-control studies, but that does not render their opinions unreliable under *Daubert*.” *Id.* at 168. In another talcum powder case, plaintiffs relied on a case-control study in which the researchers lacked an independent source of data on the study subjects' talcum powder use. But the researchers relied on a study of a different exposure, “in which retrospective recall could be compared to verifiable prospective data,” to provide a surrogate baseline measurement of inaccurate recall. *Carl v. Johnson & Johnson*, 237 A.3d 308, 327 (N.J. Super. Ct. App. Div. 2020). The researchers performed a sensitivity analysis that showed that even if their study suffered from twice as much recall bias, their result would still be statistically significant. *Id.* The court reversed the trial judge's exclusion of plaintiffs' expert testimony. *Id.* at 344.

expected to cause the disease may shed light on whether bias affected the recall of the relevant exposures.¹⁴⁵

Bias may also result from reliance on interviews with surrogates who are individuals other than the study subjects. This is often necessary when, for example, a subject (in a case-control study) has died of the disease under investigation or may be too ill to be interviewed.

The quality of a study's assessment of participants' exposure can affect the credibility of an association with a health outcome.¹⁴⁶ The extent of exposure¹⁴⁷ can be obtained from a number of different sources, each of which has limitations and strengths.¹⁴⁸

For estimating past exposures, epidemiologists often use indirect measures of exposure, such as interviewing workers and reviewing employment records. For

145. Analogously, an expert testified that because recall bias would be expected to affect reported exposures for people with any type of cancer, concerns about recall bias were diminished where "epidemiology studies on the whole observed associations only between" exposure to an herbicide and a particular type of cancer, rather than with "the other cancers about which participants were asked." *In re Roundup Prods. Liab. Litig.*, 390 F. Supp. 3d 1102, 1121 (N.D. Cal. 2018). The court concluded that "concerns about recall bias in these studies do not demand that a reliable expert opinion meaningfully discount the body of case-control studies when assessing causation." *Id.*

146. The quality of exposure assessment in an epidemiologic study should be distinguished from the issue of the quality of the evidence of the extent of the plaintiff's exposure, an issue that often arises in environmental-exposure cases. *See infra* note 150.

147. The exposure "dose" is defined in various ways, but dose often refers to the intensity or magnitude of exposure multiplied by the time exposed. *See* Bernard D. Goldstein, *Toxic Torts: The Devil Is in the Dose*, 15 J.L. & Pol'y 551, 554 (2008) ("Dose is defined as concentration multiplied by frequency or duration—it is not just the exposure level at any one point in time."). Other definitions of dose may be more appropriate in light of the biological mechanism of the disease. *See* Emily White et al., *Principles of Exposure Measurement in Epidemiology: Collecting, Evaluating, and Improving Measures of Disease Risk Factors* (2d ed. 2008); National Research Council, *Exposure Science in the 21st Century* (2012). *See also* *Gates v. Rohm & Haas Co.*, 655 F.3d 255, 260 (3d Cir. 2011) (affirming district court decision that an expert report that assumed a constant and average value of air pollutant exposure during lengthy time periods, which is considered "overly simplistic," was insufficient evidence to establish dose). For a discussion of the difficulties of determining dose from atomic fallout, *see Allen*, 588 F. Supp. 247, 425–26, *rev'd on other grounds*, 816 F.2d 1417 (10th Cir. 1987).

The timing of exposure in relation to the life course may also be critical, especially if the disease of interest is a birth defect. In *Smith v. Ortho Pharm. Corp.*, 770 F. Supp. 1561, 1577 (N.D. Ga. 1991), the court criticized a study for its inadequate measure of exposure to spermicides. The researchers had defined exposure as receipt of a prescription for spermicide within 600 days of delivery, but this definition of exposure is too broad because environmental agents are likely to cause birth defects only during a narrow band of time.

148. *See In re Paoli R.R. Yard PCB Litig.*, No. 86-2229, 1992 U.S. Dist. LEXIS 18430, at *9–*11 (E.D. Pa. Oct. 21, 1992) (discussing valid methods of determining exposure to chemicals). Exposure science is a discipline that assesses exposure and its extent and is addressed in another reference guide in this manual. *See* M. Elizabeth Marder & Joseph V. Rodricks, *Reference Guide on Exposure Science and Exposure Assessment*, "Understanding Human Exposure: Key Concepts," in this manual.

example, a study might treat all those employed to install asbestos insulation as having been exposed to asbestos during the period that they were so employed. However, there may be a wide variation of exposure within any job, and therefore this job-classification measure may not accurately determine individual levels of exposure.¹⁴⁹ If the agent of interest is a drug, medical or hospital records often can be used to determine past exposure. Retrospective studies that determine past exposures indirectly are usually less accurate than prospective studies or follow-up studies, including ones in which a drug or medical intervention is the agent of interest.

Exposure to the agent can also be measured directly. To directly determine an individual's level, duration, and timing of exposure to an agent, researchers can use biological markers, or biomarkers, when they are available.¹⁵⁰ A biomarker

149. Occupational epidemiologists frequently employ study designs that consider all agents to which workers in a particular occupation are exposed because they seek to determine the hazards associated with that occupation. Isolating one of the agents for examination would be difficult if not impossible. These studies, then, present difficulties when used in court in support of a claim by a plaintiff who was exposed to only one or some of the agents that were present at the worksite that was the subject of the study. See, e.g., *Knight v. Kirby Inland Marine Inc.*, 482 F.3d 347, 352–53 (5th Cir. 2007) (concluding that case-control studies of cancer that entailed exposure to a variety of organic solvents at job sites did not support claims of plaintiffs who claimed exposure to benzene caused their cancers).

150. A different, but related, problem often arises in court. Determining the plaintiff's exposure to the alleged toxic substance always involves a retrospective determination and may involve difficulties similar to those faced by an epidemiologist planning a study. Thus, in *John's Heating Serv. v. Lamb*, 46 P.3d 1024 (Alaska 2002), plaintiffs were exposed to carbon monoxide because of defendants' negligence with respect to a home furnace. The court observed: "[W]hile precise information concerning the exposure necessary to cause specific harm to humans and exact details pertaining to the plaintiff's exposure are beneficial, such evidence is not always available, or necessary, to demonstrate that a substance is toxic to humans given substantial exposure and need not invariably provide the basis for an expert's opinion on causation." *Id.* at 1035 (quoting *Westberry v. Gislaved Gummi AB*, 178 F.3d 257, 264 (4th Cir. 1999)). See generally Restatement (Third) of Torts: Liability for Physical and Emotional Harm § 28 cmt. c(2) & rptrs. note (2010).

Even if a biomarker exists for exposure to a particular agent, it may be problematic to attempt to use that biomarker as a measure of a plaintiff's past exposure to that agent. As described in the next paragraph, the persistence of biomarkers in the body following an exposure varies. Because of the long latency period between exposure and the manifestation of disease that is characteristic of many toxic tort cases, biomarker measurements that are available at the time of litigation may not accurately reflect a plaintiff's much earlier exposure to a toxicant.

To address the problem of determining a plaintiff's exposure in asbestos cases, a number of courts have adopted a requirement that the plaintiff demonstrate (1) regular use by an employer of the defendant's asbestos-containing product; (2) the plaintiff's proximity to that product; and (3) exposure over an extended period of time. See, e.g., *Lohrmann v. Pittsburgh Corning Corp.*, 782 F.2d 1156, 1162–64 (4th Cir. 1986); *Gregg v. V-J Auto Parts, Inc.*, 943 A.2d 216, 226 (Pa. 2007); see also *James v. Bessemer Processing Co.*, 714 A.2d 898, 911–14 (N.J. 1998) (applying same standard to a claim involving substances other than asbestos and holding that plaintiff presented sufficient evidence of exposure to certain substances in particular manufacturers' products to withstand summary judgment). Other courts have imposed more stringent requirements. E.g., *Bostic v. Georgia-Pacific Corp.*, 439 S.W.3d 332, 353 (Tex. 2014) (holding that "the dose must be quantified but need not be established with mathematical precision," although "the plaintiff must prove with

may be the actual parent compound to which an individual is exposed, a metabolite of the parent compound, or some other measure that reflects an exposure-induced alteration in biologic systems at the functional or molecular level.¹⁵¹ Measurements can be made in various biologic materials, including blood, urine, amniotic fluid, breast milk, seminal fluid, fecal material, hair, and nails. The appropriate matrix for measuring exposure is determined by the chemical agent and its properties.

Measurements in bodily fluids of chemicals with long half-lives (in the order of years), such as lead or DDT, are more likely to reflect exposure preceding a health outcome.¹⁵² Other chemical agents have short half-lives in the human body (e.g., bisphenol A, phenols), and thus the level of exposure indicated by a biomarker may not accurately reflect exposure during a critical window but may only reflect a very recent exposure (e.g., the last meal).¹⁵³ Appropriate biological markers, however, are available only for a small number of toxicants.¹⁵⁴

scientifically reliable expert testimony that the plaintiff's exposure to the defendant's product more than doubled the plaintiff's risk of contracting the disease"). For criticism of Bostic's requirement of a risk-doubling dose from each defendant's product, see Steve C. Gold, *Drywall Mud and Muddy Doctrine: How Not to Decide a Multiple-Exposure Mesothelioma Case*, 49 Ind. L. Rev. 117 (2015).

151. Courts occasionally have noted the use of biomarkers to classify study subjects' exposure levels. See, e.g., *Nat'l Bank of Commerce (of El Dorado, Ark.) v. Dow Chem. Co.*, 965 F. Supp. 1490, 1553–54 (E.D. Ark. 1996) (favorably citing defendant's expert's testimony that praised a study in which low levels of cholinesterase in blood were used as a biomarker for exposure to cholinesterase-inhibiting pesticides). More often, courts must resolve disputes about whether a particular plaintiff displays a putative biomarker of exposure and the significance of the putative biomarker's presence or absence. See, e.g., *In re TMI Litig.*, 193 F.3d 613, 690–93 (3d Cir. 1999) (affirming exclusion of expert's estimates of radiation doses plaintiffs received during nuclear accident fifteen years earlier because another methodology would provide more accurate estimates after such a long lapse of time); *In re Flint Water Cases*, 2021 WL 5356295, at *3 (E.D. Mich. Nov. 17, 2021) (denying motion to exclude testimony of plaintiff's expert that measurement of lead levels in bone, rather than in blood, provided more accurate exposure assessment); *Barlow v. Gen. Motors Corp.*, 595 F. Supp. 2d 929, 944 n.6 (S.D. Ind. 2009) (rejecting plaintiffs' suggestion that body fat measurements would provide better marker of PCB exposure than blood measurements, where plaintiffs failed to produce evidence to that effect); *Young v. Burton*, 567 F. Supp. 2d 121, 136 (D.D.C. 2008) (excluding plaintiff's expert testimony that certain biomarkers indicated exposure to mold five years earlier, where defense experts testified that the markers' response to exposure lasts only minutes or hours).

152. If such a biomarker exists, retrospective studies may be able accurately to capture the exposure during critical biological windows depending on the half-life of the chemical and the time since that window.

153. In cases where the exposure has a short half-life, multiple measurements may be needed to accurately assess the level of exposure, unless they are reflective of typical exposure or if a single exposure can have a biologic effect. To use a biomarker to obtain an accurate assessment of exposure to such an agent, it may be necessary to measure the biomarker at multiple times, unless it is shown that a single measurement is representative of a longer-term exposure or that a single exposure could cause the health outcome at issue.

154. See Jessie P. Buckley et al., *Exposure to Contemporary and Emerging Chemicals in Commerce Among Pregnant Women in the United States: The Environmental Influences on Child Health Outcome*

In addition to biomarkers, epidemiologists may use measurements of non-biologic matrices such as water, dust, and air to assess exposure to agents found in those media.¹⁵⁵ Sometimes monitoring devices can be used to measure such environmental exposures directly, but monitoring data often are not available for exposures that have occurred in the past unless they are routinely collected, such as through governmental air or water monitoring.

The route (e.g., inhalation or absorption), duration, and intensity of exposure also are important factors. Even with environmental monitoring, the dose measured in the environment generally is not the same as the dose that reaches internal target organs. If the researcher has calculated the internal dose of exposure, the scientific basis for this calculation should be examined for soundness.¹⁵⁶

Disease information bias

Information bias may also result from inaccurate measurement of disease status. The quality and sophistication of the diagnostic methods used to detect a disease should be assessed.¹⁵⁷ The proportion of subjects eligible for study who were

(ECHO) Program, 56 Env't Sci. & Tech. 6560, 6560 (2022), <https://doi.org/10.1021/acs.est.1c08942> (National Health and Nutrition Examination Survey (NHANES) provides data on exposure to 350 chemicals, a small fraction of the number of chemicals in use); Angela N.H. Creager, *Human Bodies as Chemical Sensors: A History of Biomonitoring for Environmental Health and Regulation*, 70 Stud. Hist. & Phil. Sci. 70, 76 (2018), <https://doi.org/10.1016/j.shpsa.2018.05.010> ("For the past quarter century, research on biomarkers and genetic dosimetry has seemed poised to change environmental public health by offering more sensitive tools for measuring chemical exposure. Yet so far these tools have not displaced the widespread reliance on methods of chemical analysis."); Steve C. Gold, *The More We Know, The Less Intelligent We Are?—How Genomic Information Should, And Should Not, Change Toxic Tort Causation Doctrine*, 34 Harv. Env't L. Rev. 370 (2010) (describing circumstances where courts can improve their ability to link toxic substances to human illness by properly understanding and utilizing genomic information and analyzing biomarkers); Gary E. Marchant, *Genetic Data in Toxic Tort Litigation*, 14 J. L. & Pol'y 7, 18–32 (2006) (explaining concept of biomarkers and how they might be used to provide evidence of: (1) exposure or dose, (2) general and specific causation, and (3) identifying those with greater risks of future disease; discussing cases in which biomarkers were employed; and concluding, "genomic data [like biomarkers] have enormous potential to make toxic tort litigation more informed, consistent and fair").

155. For example, concentrations of perfluorooctanoic acid (PFOA, a so-called "forever chemical") found in a drinking water supply, levels of lead in dust in a home, or levels of polycyclic aromatic hydrocarbons (PAHs) or fine particulate matter (PM_{2.5}) in the ambient air may be used to estimate exposures of the people who drink the water, live in the home, or breathe the air.

156. See also Eaton et al., *supra* note 60.

157. The hazards of adversarial review of epidemiologic studies to determine bias are highlighted by *O'Neill v. Novartis Consumer Health, Inc.*, 55 Cal. Rptr. 3d 551, 558–60 (Ct. App. 2007). Defendant's experts criticized a case-control study relied on by plaintiff on the ground that there was misclassification of exposure status among the cases. Plaintiff objected to this criticism because defendant's experts had only examined the cases for exposure misclassification, which would tend

actually examined should be determined. If, for example, many of the subjects refused to be tested, the fact that the test used was of high quality may be of relatively little value given the potential for selection bias.

The scientific validity of the research findings is influenced by the accuracy of the diagnosis of disease or health status under study.¹⁵⁸ The disease must be one that is recognized and defined to enable accurate diagnoses. To ensure that the diagnoses of study subjects are in fact accurate, diagnostic criteria accepted by the medical community should be used.¹⁵⁹ For example, a researcher

to exaggerate any association by providing an inaccurately inflated measure of exposure in the cases. The experts failed to examine whether there was misclassification in the controls, which, if it existed, would tend to incorrectly diminish any association.

158. In *In re Swine Flu Immunization Prods. Liab. Litig.*, 508 F. Supp. 897, 903 (D. Colo. 1981), *aff'd sub nom. Lima v. United States*, 708 F.2d 502 (10th Cir. 1983), the court critically evaluated a study relied on by an expert whose testimony was stricken. In that study, determination of whether a patient had Guillain-Barré syndrome was made by medical clerks, not physicians who were familiar with diagnostic criteria.

159. See generally Robert H. Friis & Thomas A. Sellers, *Epidemiology for Public Health Practice* 494 (6th ed. 2021) (in a case-control study, “the definition of a case is influenced by a number of factors, including whether there are standard diagnostic criteria . . . and whether the criteria to diagnose the disease are subjective or objective”).

Classification problems have affected the study of agent-disease links that later became the subject of litigation, for example the under-reporting of deaths due to mesothelioma. See *supra* note 140. Few court opinions, however, have addressed misclassification bias as the basis of a challenge to the validity of epidemiologic studies. See, e.g., *In re Lipitor (Atorvastatin Calcium) Mktg., Sales Pracs., & Prods. Liab. Litig.* (No. II) MDL 2502, 892 F.3d 624, 635–38 (4th Cir. 2018) (affirming exclusion of expert testimony where expert based opinion on reanalysis of epidemiologic study data using a different disease definition); *Rochkind v. Stevenson*, 164 A.3d 254, 261–62 (Md. 2017) (reversing admission of expert testimony opining that exposure to lead can cause attention deficit-hyperactivity disorder (ADHD) where expert relied on epidemiologic studies showing association with attention decrements in general but not with ADHD as defined by diagnostic criteria); *In re Lipitor (Atorvastatin Calcium) Mktg., Sales Pracs., & Prods. Liab. Litig.*, 174 F. Supp. 3d 911, 917 n.5 (D.S.C. 2016) (noting that where one study found a statistically significant association, but another did not, the difference could be explained by the latter study’s use of a more restrictive disease definition but making “no value judgments” with respect to either study’s quality).

A related but subtly different issue arises more frequently: disputes about whether a plaintiff has been diagnosed accurately with a disease that has been (or could be) the subject of epidemiologic study. See, e.g., *Grant v. Bristol-Myers Squibb*, 97 F. Supp. 2d 986, 992 (D. Ariz. 2000) (“where experts propose that breast implants cause a disease but cannot specify the criteria for diagnosing the disease, it is incapable of epidemiologic testing[, rendering] the experts’ methods insufficiently reliable to help the jury”); *Burton v. Wyeth-Ayerst Labs.*, 513 F. Supp. 2d 719, 722–24 (N.D. Tex. 2007) (parties disputed whether cardiology problem involved two separate diseases or only one; court concluded that all experts in the case reflected a view that there was but a single disease); *Nat’l Bank of Commerce (of El Dorado, Ark.) v. Dow Chem. Co.*, 965 F. Supp. 1490 (E.D. Ark. 1996) (where children exposed *in utero* to pesticides suffered from birth defects, parties disputed whether children’s symptoms supported diagnosis of inherited syndromes); *In re Breast Implant Cases*, 942 F. Supp. 958, 961 (E.D.N.Y. & S.D.N.Y. 1996) (rejecting plaintiffs’ experts’ testimony that breast implants caused a “new undifferentiated atypical disease” with “hundreds of symptoms”).

interested in studying spontaneous abortion in the first trimester must determine that study subjects are pregnant. If that determination were made based on the results of home pregnancy tests at a time when home pregnancy tests were known to have a high rate of false-positive results (indicating pregnancy when the woman is not pregnant), the study will overestimate the number of spontaneous abortions.

Information bias due to misclassification

Misclassification is a consequence of information bias in which, because of problems with the information available, individuals in the study may be misclassified as to exposure status or disease status. Bias due to exposure misclassification can be differential or nondifferential. In nondifferential misclassification (also called random misclassification), the inaccuracies in determining exposure are independent of disease status, or the inaccuracies in diagnoses are independent of exposure status—in other words, the data are crude and contain some level of random error. This is a common problem.

Generally, nondifferential misclassification shifts the study's result toward a finding of no effect.¹⁶⁰ Thus, if errors are nondifferential, an apparent association between an exposure and disease would not be a product of incorrect classification. Instead, nondifferential misclassification generally underestimates the true size of the association. Studies that find no association and are likely affected by a fair amount of nondifferential misclassification should therefore be regarded with suspicion. In fact, it has been shown that under some conditions, severe nondifferential misclassification could result in an association that suggests that an agent reduces the risk of disease, when in fact exposure to the agent increases the risk. However, it is important to note that nondifferential misclassification on its own does not guarantee a bias towards the null. Bias away from the null, resulting in a larger estimated association than the truth, can occur if the exposure or the health outcome has more than two levels (i.e., present versus absent)¹⁶¹ or if misclassification is related to errors made on other variables.¹⁶²

160. Toxic tort cases are typically concerned with the association of an exposure with disease, a dichotomous outcome that is either present or absent. In studies of dichotomous outcomes, nondifferential misclassification generally shifts the observed value of the relative risk, odds ratio, or other measure of association toward one. In studies between exposures and continuous variables such as IQ, which measure associations differently, nondifferential misclassification generally shifts the observed value of the study's measure of association toward zero.

161. See Alexander Walker et al., *Comparing Imperfect Measures of Exposure*, 121 Am. J. Epidemiology 783 (1985), <https://doi.org/10.1093/oxfordjournals.aje.a114049>; Mustafa Dosemeci et al., *Does Nondifferential Misclassification of Exposure Always Bias a True Effect Toward the Null Value?*, 132 Am. J. Epidemiology 746 (1990), <https://doi.org/10.1093/oxfordjournals.aje.a115716>.

162. See Michael Chavance et al., *Correlated Nondifferential Misclassifications of Disease and Exposure: Application to a Cross-Sectional Study of the Relation Between Handedness and Immune Disorders*, 21

Differential misclassification is caused by a systematic error in determining exposure in cases as compared with controls, or disease status in unexposed cohorts relative to exposed cohorts. For example, in a case-control study this would occur if, in the process of anguishing over the possible causes of the disease, parents of ill children recalled more exposures to a particular agent than actually occurred, or if parents of the controls, for whom the issue was less emotionally charged, recalled fewer. This can also occur in a cohort study in which, for example, birth-control users (the exposed individuals) are monitored more closely for potential side effects, leading to a higher rate of disease identification in that cohort than in the unexposed cohort. Depending on how the misclassification occurs, a differential bias can produce an error in either direction—the exaggeration or understatement of a true association.

Confounding Bias

Confounding is another important potential source of error in epidemiologic studies.¹⁶³ Confounding occurs when another causal factor (the confounder) confuses the relationship between the agent and the outcome of interest.¹⁶⁴ Confounding can bias a study result by either exaggerating or diluting an association.¹⁶⁵ One instance of confounding is when a factor is causally related to both the agent and disease of interest. For example, researchers may conduct a study that finds individuals with gray hair have a higher rate of death than those with hair of another color. Instead of hair color having an effect on death, the results might be explained by the confounding factor of age. If old age causes gray hair (those who are older have a greater probability of having gray hair than those who are younger), old age may be responsible for the association found between hair color and death.¹⁶⁶ To negate the effect of confounding bias, researchers

Int'l J. Epidemiology 537 (1992), <https://doi.org/10.1093/ije/21.3.537>; P. Kristensen, *Bias from Nondifferential but Dependent Misclassification of Exposure and Outcome*, 3 Epidemiology 210 (1992), <https://doi.org/10.1097/00001648-199205000-00005>.

163. See *In re Abilify (Aripiprazole) Prods. Liab. Litig.*, 299 F. Supp. 3d 1291, 1322–26 (N.D. Fla. 2018) (quoting this reference guide and discussing the evidence that negated the possibility of confounding in an epidemiologic study).

164. See Lash et al., *supra* note 50, at 263–64.

165. One example of a confounding factor that may result in a study's outcome understating an association is vaccination. Thus, if a group exposed to an agent has a higher rate of vaccination for the disease under study than the unexposed group, the vaccination may reduce the rate of disease in the exposed group, thereby producing an association that is less than the true association without the confounding of vaccination.

166. This example is drawn from Kahn & Sempos, *supra* note 34, at 63.

must separate the relationship between gray hair and risk of death from that of old age and risk of death.¹⁶⁷

Confounding can be further illustrated by a hypothetical prospective cohort study designed to investigate whether drinking alcohol is associated with emphysema. Participants are followed for a period of 20 years and the incidence of emphysema in the exposed (participants who consume more than 15 drinks per week) and the unexposed is compared. At the conclusion of the study, the relative risk of emphysema in the drinking group is found to be 2.0 (an association that suggests a possible effect). But does this association reflect a true causal relationship or might it be the product of confounding?

One possibility for a confounding factor is smoking, a known causal risk factor for emphysema. If those who drink alcohol are more likely to be smokers than those who do not drink, then smoking may be responsible for some or all of the higher risk of emphysema among those who drink.

A serious problem in observational studies such as this hypothetical study is that the individuals are not assigned randomly to the groups being compared as they might be in a clinical trial. As discussed above, randomization maximizes the possibility that factors other than the one under study are evenly distributed between the exposed and the unexposed groups.¹⁶⁸ In observational studies, by contrast, other forces determine who is exposed to other (possibly causal) factors. The lack of randomization leads to the potential problem of confounding. Thus, for example, the exposed group might consist of those who are exposed at work to an agent suspected of being an industrial toxicant. The members of this group may, however, differ from unexposed controls by residence, socioeconomic or health status, age, or other extraneous factors.¹⁶⁹ These other factors may be causing (or protecting against) the disease, but because of potential confounding,

167. *Schwab v. Philip Morris USA, Inc.*, 449 F. Supp. 2d 992, 1199–1200 (E.D.N.Y. 2006), *rev'd on other grounds*, 522 F.3d 215 (2d Cir. 2008), describes confounding that led to premature conclusions that low-tar cigarettes were safer than regular cigarettes. Smokers who chose to switch to low-tar cigarettes were different from other smokers in that they were more health conscious in other aspects of their lifestyles. Failure to account for that confounding—and measuring a healthy lifestyle is difficult even if it is identified as a potential confounder—biased the results of those studies.

168. Randomization attempts to ensure that the presence of a potentially confounding risk factor (such as smoking in the preceding paragraph's example), is governed by chance, as opposed to being affected or determined by the presence of the potential risk factor under study (such as drinking in the preceding paragraph's example) or the underlying medical condition being studied (such as emphysema in the preceding paragraph). See Kenneth J. Rothman & Timothy L. Lash, *Epidemiologic Study Design with Validity and Efficiency Considerations*, in Lash et al., *supra* note 50, at 105, 109, 121–22; see also section titled “Experimental and Observational Studies of Suspected Toxic Agents” above.

169. See, e.g., *In re “Agent Orange” Prod. Liab. Litig.*, 597 F. Supp. 740, 783 (E.D.N.Y. 1984) (discussing the problem of confounding that might result in a study of the effect of exposure to Agent Orange on Vietnam servicemen), *aff'd*, 818 F.2d 145 (2d Cir. 1987).

an association of the disease with exposure to the agent may appear when one does not exist or may be masked when one exists. Confounders, like smoking in the alcohol-drinking study, do not reflect an error made by the investigators; rather, they reflect the inherently “uncontrolled” nature of exposure designations in observational studies. When they can be identified, confounders should be taken into account.

To evaluate whether smoking is a confounding factor in a study of alcohol consumption and emphysema, the researcher would stratify the data into smoking and nonsmoking subgroups, would compute associations within each subgroup, and would average these results. If the average of the stratified associations was the same as that in the all-subjects group, then smoking would not be a confounding factor. But if these values differed, this would provide evidence that smoking is a confounder of the observed association between drinking and emphysema. If the stratified results showed no association (e.g., a relative risk of 1), this would indicate that the confounding factor (smoking) fully accounts for the apparent association of drinking with emphysema.

Table 4 reveals our hypothetical study’s results, with smoking being a confounding factor, which, when accounted for, eliminates the association. As the table shows, in the full cohort, drinkers have twice the risk of emphysema compared with nondrinkers. When the relationship between drinking and emphysema is examined separately in smokers and in nonsmokers, the risk of emphysema in drinkers compared with nondrinkers is not elevated in smokers or in nonsmokers. This is because smokers are disproportionately drinkers and have a higher rate of emphysema than nonsmokers. Thus, the relationship between drinking and emphysema in the full cohort is distorted by failing to take into account the relationship between being a drinker and being a smoker.

Even after accounting for the effect of known suspected confounders (such as smoking in our example), there is always a risk that an undiscovered or unrecognized confounding factor may contribute to a study’s findings, by either magnifying or reducing the observed association.¹⁷⁰ It is, however, necessary to keep that risk in perspective. Often the mere possibility of uncontrolled confounding is used to call into question the results of a study. This was certainly the strategy of some who sought, or unwittingly helped, to undermine the implications of the studies persuasively linking cigarette smoking to lung cancer. The critical question is whether it is plausible that the findings of a given study could indeed be due to unrecognized confounders.

170. Tyler J. VanderWeele & Kenneth J. Rothman, *Formal Causal Models*, in Lash et al., *supra* note 50, at 33; *see also* section titled “Experimental and Observational Studies of Suspected Toxic Agents” above.

Table 4. Hypothetical Emphysema Study Data^a

	Total Cohort				Smokers				Nonsmokers			
	Total	Cases	Incidence ^a	RR ^b	Total	Cases	Incidence ^a	RR	Total	Cases	Incidence ^a	RR
Drinking Status												
Nondrinkers	471	16	0.034	—	111	9	0.081	—	360	7	0.019	—
Drinkers	739	51	0.069	2.0	592	48	0.081	1.0	147	3	0.020	1.0

^a The incidence of disease is not normally presented in an epidemiologic study, but we include it here to aid in comprehension of the ideas discussed in the text.

^b RR = relative risk. The relative risk for the drinkers in each of the cohorts is determined based on reference to the risk among nondrinkers; that is, the risk of disease among drinkers is compared with nondrinkers for each of the three cohorts separately.

Techniques to identify confounding factors

Epidemiologic theory has evolved over recent years and so has the understanding of confounding. Directed acyclic graphs (DAGs), a form of causal diagram, are now considered to be the tool of choice to identify which factors should be addressed as potential confounders. These graphs reflect the understanding of the researcher regarding the causal relation among factors (shown as nodes) by linking them with arrows pointing from causes to effects. Any path from exposure to the health outcome under study that does not follow the direction of arrows can generate statistical associations that are not causal. Examination of this graph will thus allow the researcher to determine which factors to adjust for to block noncausal associations.

Figure 6 illustrates how epidemiologists can use DAGs to identify confounders. For the hypothetical study depicted in Figure 6, the research question is whether exposure A to the herbicide glyphosate increases the risk of outcome Y, non-Hodgkin lymphoma. But suppose some potential confounding factors C, such as age, race, and sex, affect both the exposure A and the outcome Y. Focusing for simplicity on just one of these factors, age, suppose that older people had higher exposure to glyphosate and also had higher risk of non-Hodgkin lymphoma. This would create a noncausal association between exposure and disease. The DAG in Figure 6 reveals that a noncausal path—one that does not follow the causal direction indicated by arrows—can be drawn between exposure A and outcome Y by passing through confounder C. If this hypothetical study finds an association between exposure to glyphosate and non-Hodgkin lymphoma risk, at least part of that observed association could be due to the fact that older people, because of their age, experience both higher exposure and greater risk of the disease.

DAGs can also allow investigators to determine which factors should not be adjusted for in order to avoid blocking a causal path.

Figure 6. Directed acyclic graph (DAG) for a hypothetical study of glyphosate and non-Hodgkin lymphoma.

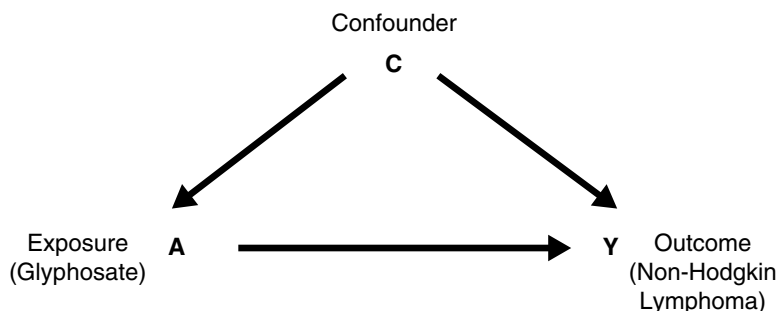
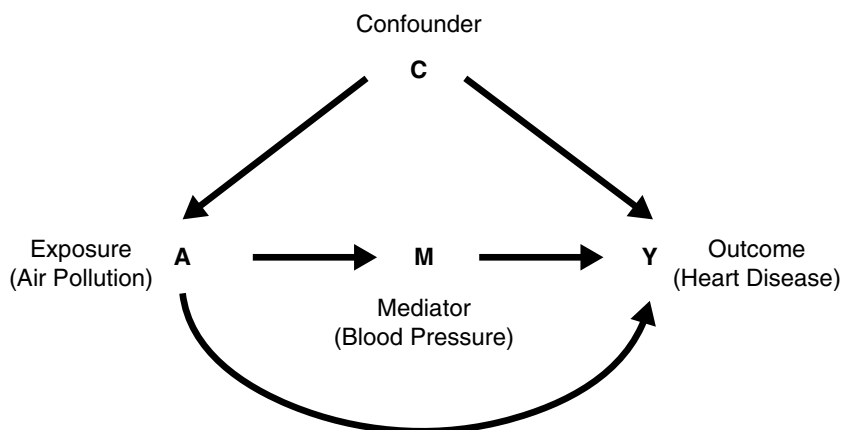


Figure 7. Directed acyclic graph for a hypothetical study of air pollution and heart disease.



For instance, imagine a hypothetical study of a possible association between exposure to air pollution and risk of heart disease. Based on a review of prior research, the epidemiologist suspects that air pollution may cause heart disease, at least in part, by affecting blood pressure. In this causal model, depicted in a DAG in Figure 7, blood pressure is assumed to mediate, rather than to confound, any causal connection between air pollution and heart disease. The DAG shows that it would be inappropriate to adjust for blood pressure, as this would block the portion of the effect of air pollution that occurs through blood pressure. This example illustrates that adjusting for factors can either remove or introduce bias. Careful examination of the causal relation among factors is necessary to determine the appropriate approach to apply.

Techniques to prevent or limit confounding

Choices in the design of a research project (e.g., methods for selecting the subjects) can prevent or limit confounding. In designing a study, the researcher must determine other risk factors for the disease under study, such as age, sex, or genotype, that may be related to exposure to the agent under study. Investigators can limit the differential distribution of these factors in the study groups by selecting unexposed subjects to “match” exposed subjects in terms of these variables. If the two groups are matched (for example, by age), then any association observed in the study cannot be the result of age, the matched variable.¹⁷¹ Often

171. Selecting a control population based on matched variables necessarily affects the representativeness of the selected controls and may affect how generalizable the study results are to the

investigators frequency-match groups so that the distribution (for example, of age) is similar in the two groups. This reduces the potential for confounding by age but may still require adjustment.

Restricting the persons who are included as subjects in a study is another method to control for confounders. If age or sex is suspected as a confounder, then the subjects enrolled in a study can be limited to those of one sex or those who are within a specified age range. When there is no variance among subjects in a study with regard to a potential confounder, confounding as a result of that variable is eliminated.

In most observational studies, many factors could potentially confound associations. This makes it challenging to apply the matching or restriction methods. Indeed, if age, socioeconomic status, smoking, and race are identified as potential confounders, finding matches or restricting participation to individuals with very precise characteristics may be impractical and make it difficult to generalize findings to individuals with other characteristics than those included in the study. For this reason, most studies use statistical models to control for multiple confounders.

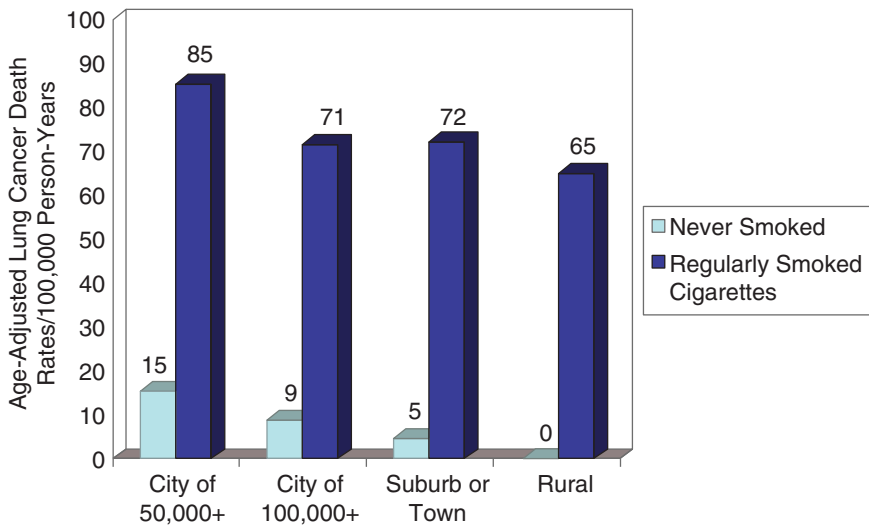
Techniques to control for confounding factors

A good study design will consider potential confounders and obtain data about these variables. There are several analytic approaches to account for the distorting effects of a potential confounder, including stratification or multivariable regression analysis. Stratification permits an investigator to evaluate the effect of a suspected confounder by subdividing the study groups based on a confounding factor. For example, in Table 4, alcohol drinkers have been stratified based on whether they smoke (the suspected confounder).

To take another example that entails a continuous rather than dichotomous potential confounder, let us say we are interested in the relationship between smoking and lung cancer but suspect that air pollution or urbanization may confound the relationship. An observed relationship between smoking and lung cancer could then theoretically be due in part to pollution, if smoking were more common in polluted areas. One could investigate this issue by stratifying the data by degree of urbanization and looking at the relationship between smoking and lung cancer in each urbanization stratum. Figure 8 shows actual

population at large. However, for a study to have merit, it must first be internally valid; that is, it must not be subject to unreasonable sources of bias or confounding. Only after a study has been shown to meet this standard does its universal applicability or generalizability to the population at large become an issue. When a study population is not representative of the general or target population, existing scientific knowledge may permit reasonable inferences about the study's broader applicability, or additional confirmatory studies of other populations may be necessary.

Figure 8. Age-adjusted lung-cancer mortality rates per 100,000 person-years by urban-rural classification and smoking category.



Source: Adapted from E. Cuyler Hammond & Daniel Horn, *Smoking and Death Rates: Report on Forty-Four Months of Follow-Up of 187,783 Men: II, Death Rates by Cause*, 166 JAMA 1294 (1958).

age-adjusted lung-cancer mortality rates per 100,000 person-years by urban or rural classification and smoking category.¹⁷²

For each degree of urbanization, lung cancer mortality rates in smokers are shown by the darker bars, and nonsmoker mortality rates are indicated by lighter bars. From these data we see that in every level (or stratum) of urbanization, lung cancer mortality is higher in smokers than in nonsmokers. Therefore, the observed association of smoking and lung cancer cannot be attributed to the level of urbanization. By examining each stratum separately, we, in effect, hold urbanization constant and still find much higher lung cancer mortality in smokers than in nonsmokers.

The standardization method was historically used primarily to control for the effect of age in mortality studies. Two types of standardizations are used—direct and indirect. In direct standardization (e.g., when based on age), overall disease or death rates are calculated for each population as though each had the age distribution of another standard or reference population, using the age-specific disease or death rates for each study population. We can then compare these overall rates, called age-adjusted rates, knowing that any difference

172. This example and Figure 8 (including the source credit in Figure 8) are from Celentano & Szklo, *supra* note 37, at 294–99.

between these rates cannot be attributed to differences in age, since both age-adjusted rates were generated using the same standard population.

Indirect standardization is used when the age-specific rates for a study population are not known. In that case, the overall disease/death rate for the standard/reference population is recalculated based on the age distribution of the population of interest using the age-specific rates of the standard population. Then, the actual number of disease cases/deaths in the population of interest can be compared with the number in the reference population that would be expected if the reference population had the age distribution of the population of interest.

This ratio is called the standardized mortality ratio (SMR). When the outcome of interest is disease rather than death, it is called the standardized morbidity ratio¹⁷³ (and the discussion below of standardized mortality rate is equally applicable to a standardized morbidity rate). If the ratio equals 1.0, the observed number of deaths equals the expected number of deaths, and the mortality rate of the population of interest is no different from that of the reference population. If the SMR is greater than 1.0, the population of interest has a higher mortality risk than that of the reference population, and if the SMR is less than 1.0, the population of interest has a lower mortality rate than that of the reference population.

Thus, age adjustment provides a way to compare populations while in effect holding age constant. Adjustment is used not only for comparing mortality rates in different populations but also for comparing rates in different groups of subjects selected for study in epidemiologic investigations. Although this discussion has focused on adjusting for age, it is also possible to adjust for other variables, such as sex, race, genetic variability, occupation, and socioeconomic status.¹⁷⁴ However, because this method is based on stratifying study data into groups that will become smaller as the number of variables considered increases, the standardization method can only be used to account for a small number of potential confounders without reducing the numbers in each category to the point at which they are not statistically stable.

As stated above, most studies use statistical models to address multiple confounders at once. Multivariable regression analysis, a technique that relates an outcome to multiple explanatory factors, is most commonly used for this

173. See *Burst v. Shell Oil Co.*, No. CIV.A. 14-109, 2015 WL 3755953, at *6 n.15 (E.D. La. June 16, 2015) (explaining SMR and its relationship with relative risk), *aff'd*, 650 F. App'x 170 (5th Cir. 2016) (quoting *Taylor v. Airco, Inc.*, 494 F. Supp. 2d 21, 25 n.4 (D. Mass. 2007)). For an example of adjustment used to calculate an SMR for workers exposed to benzene, see Robert A. Rinsky et al., *Benzene and Leukemia: An Epidemiologic Risk Assessment*, 316 NEJM 1044 (1987), <http://doi.org/10.1289/ehp.8982189>.

174. For further elaboration on adjustment, see Celentano & Szklo, *supra* note 37, at 82–85; Cole, *supra* note 94, at 10,281.

purpose.¹⁷⁵ For example, Narayan and colleagues used logistic regression to estimate whether occupational exposure to pesticides was associated with the risk of Parkinson's disease in central California while controlling for multiple potential confounders including age, sex, smoking, education, race, and frequency of household pesticide use, as well as ambient residential- and work-address pesticide exposures. In this case-control study, they found that self-reported occupational use of herbicides for a duration exceeding 10 years was associated with a doubling in the risk (i.e., an odds ratio of 2.1 and 95% confidence interval of 1.1 to 3.9) of being diagnosed with Parkinson's disease.¹⁷⁶

However, because there is no free lunch, the use of such methods to address confounding comes at a cost, as they require making a number of assumptions about the characteristics of the data and of the relation among the variables.¹⁷⁷ In addition, the larger the number of confounders one wants to control for, the less precise the association between an agent and a disease generally becomes. This is referred to as the bias–variance tradeoff.

The type of multivariable regression model that is used is primarily determined by the form of the outcome data under study. For example, logistic regression (generating odds ratios) is often used for dichotomous outcome data (e.g., presence versus absence of disease); linear regression (generating differences in means) is generally used for continuous outcome data (e.g., blood pressure or IQ); and Cox proportional hazards models (generating hazard ratios) tend to be the method of choice for survival data (e.g., time to death for a given cancer).

Regression analysis was originally developed in the nineteenth century. Statistical methods have evolved, and researchers sometimes use more sophisticated approaches. For example, Bayesian analyses (named after the statistician Thomas Bayes) allow for the integration of prior knowledge in the statistical analysis.¹⁷⁸ In a combination between the exposure-matching and the regression

175. For a more complete discussion of multivariable analysis, see Rubinfeld & Card, *supra* note 31.

176. See Shilpa Narayan et al., *Occupational Pesticide Use and Parkinson's Disease in the Parkinson Environment Gene (PEG) Study*, 107 *Env't Int'l* 266 (2017), <http://doi.org/10.1016/j.envint.2017.04.010>.

177. For more details, see Rubinfeld & Card, *supra* note 31, “What Model Should Be Used to Evaluate the Question at Issue?”

178. Epidemiologic studies assessed by Bayesian statistical analyses have begun to gain a foothold in litigation, although court opinions are still dominated by discussion of traditional significance testing. See *In re Abilify (Aripiprazole) Prods. Liab. Litig.*, No. 3:16MD2734, 2021 WL 4951944, at *5 (N.D. Fla. July 15, 2021) (“Numerous federal courts have found Dr. Madigan’s methodology of detecting safety signals using a combination of frequentist and Bayesian algorithms to be reliable under Rule 702 and *Daubert*.”); *Langrell v. Union Pac. R. Co.*, No. 8:18CV57, 2020 WL 3037271, at *3 (D. Neb. June 5, 2020) (admitting testimony of specific causation expert who testified “he used a Bayesian approach” to assess causation of a cancer so rare that it was “unlikely or impossible for epidemiological studies to be performed”); *In re Testosterone Replacement Therapy Prods. Liab. Litig.*, No. 14 C 1748, 2018 WL 4030585, at *8 (N.D. Ill. Aug. 23, 2018) (denying motion to exclude testimony of expert “whose Bayesian critiques of epidemiological studies” were similar to those of another expert whose testimony “the Court has previously found admissible”). For a description of

approaches, propensity score matching allows researchers to match study subjects based on their probability of exposure given the value of potential confounders.¹⁷⁹ The potential outcome (or counterfactual) approach aims to simulate randomized-controlled trials by estimating outcomes under different exposure scenarios in which measured confounders are not related to the exposure and/or the outcome. Methods applying this approach include inverse-probability weighting and the parametric g-formula.

The methods described above, through different means and under different assumptions, allow for adjustment of the effect of confounders on an association between an agent and a disease. They aim to estimate the association after having removed confounding by the measured variables. The details of these methods are too technical for discussion here. The crucial points to recognize are that using statistical models to address possible confounding is an accepted part of epidemiologic practice and that choosing the correct statistical model for a study is important to the validity of the study's results. If an association between an agent and a disease remains after accounting for selection, information, and confounding bias, a researcher must then assess whether an inference of causation is justified. This entails consideration of the *Hill* factors explained in the section below titled "General Causation."

Conceptual Error

In addition to the possibility of bias, a study may also be limited by flawed definitions or premises in its design. Consider, for example, a researcher who wishes to investigate whether a certain drug is teratogenic (causes birth defects). If the researcher defines the disease of interest as all birth defects, rather than a specific type of birth defect, there should be a scientific basis to hypothesize that the effects of the drug could be so broad. If the effect is in fact more limited, the result of this conceptualization error could be to dilute or mask any real effect that the agent might have on a specific type of birth defect.¹⁸⁰

Bayesian statistical methods, see Kaye & Stern, *supra* note 16, "Bayesian Statistical Methods and Posterior Probabilities" and "Appendix: Conditional Probability and Bayes' Rule."

179. "Propensity scoring is a system where researchers attempt to control for confounding by creating a comorbidity 'score' for a study participant; the more confounding comorbidities a participant has, the higher [the participant's] score and, through the score, the researchers estimate the propensity of the subject to contribute to confounding." *In re Zantac* (Ranitidine) Prods. Liab. Litig., 644 F. Supp. 3d 1075, 1260 (S.D. Fla. 2022). In *Zantac*, the court excluded proffered expert testimony in part because the court found that the experts testified inconsistently about the merits of propensity scoring. *Id.* at 1261.

180. In *Brock*, 874 F.2d at 312, the court discussed a reanalysis of a study in which the effect was narrowed from all congenital malformations to limb-reduction defects. The magnitude of the association changed by 50% when the effect was defined in this narrower fashion. See Timothy J. Lash et al., *Measurement and Measurement Error*, in Lash et al., *supra* note 50, at 287, 297

Similarly, if the researcher studies the connection between the drug and limb-reduction birth defects in particular, the researcher should consider that an exogenous agent can only cause such a defect during the period of limb organogenesis (weeks 5–9 of the pregnancy). If the researcher defines exposure as the mother taking the drug anytime during her pregnancy, conceptual bias is introduced. Classifying as exposed those pregnant women who took the drug outside the window when exposure could have caused the outcome of interest would tend to understate any association that might exist.

Biases That Affect a Body of Epidemiologic Evidence

Some biases go beyond errors in individual studies and affect the overall body of available evidence in a way that skews what appears to be the universe of evidence. Publication bias is the tendency for scientific journals to prefer to publish studies that find an association.¹⁸¹ If negative studies are never published, the published literature as a whole will be biased. These types of biases may present problems for researchers who attempt to synthesize or combine the results of multiple studies to assess the existence and magnitude of an association.¹⁸²

Financial conflicts of interest by researchers and the source of funding of studies have also been shown to have an effect on the outcomes of such studies.¹⁸³ As the former editor-in-chief of the *British Medical Journal* noted:

(“Unwarranted assurances of a lack of any effect can easily emerge from studies in which a wide range of etiologically unrelated outcomes are grouped.”). For a more general discussion of conceptualization and other types of errors and their relation to legal decision-making, see Vern R. Walker, *Theories of Uncertainty: Explaining the Possible Sources of Error in Inferences*, 22 *Cardozo L. Rev.* 1523, 1544–47, 1553–55 (2001).

181. Celentano & Szklo, *supra* note 37, at 234. Investigators may contribute to this effect by neglecting to submit negative studies for publication. Publication bias “can pose a serious problem when the results from all published clinical trials are reviewed.” *Id.* Publication bias can also affect systematic reviews or meta-analyses of observational studies on a given agent–disease relationship. For clinical trials, medical journals have attempted to minimize publication bias by requiring that before any study is considered for publication, it must have been registered in a public trial registry before any data is gathered. See Catherine D. DeAngelis et al., *Clinical Trial Registration: A Statement from the International Committee of Medical Journal Editors*, 292 *JAMA* 1363 (2004), <https://doi.org/10.1056/NEJMe048225>. The National Institutes of Health (NIH) requires that all NIH-funded clinical trials must be registered and the results reported. See Nat’l Insts. of Health, *NIH Policy on the Dissemination of NIH-Funded Clinical Trial Information*, NOT-OD-16-149, Sept. 16, 2016, <https://grants.nih.gov/grants/guide/notice-files/NOT-OD-16-149.html>.

182. See section titled “Methods for Synthesizing or Combining the Results of Multiple Studies” below.

183. See Jerome P. Kassirer, *On the Take: How Medicine’s Complicity with Big Business Can Endanger Your Health* 79–84 (2005); Deepa V. Cherla et al., *The Effect of Financial Conflict of Interest, Disclosure Status, and Relevance on Medical Research from the United States*, 34 *J. Gen. Internal Med.*

The major determinant of whether reviews of passive smoking concluded it was harmful was whether the authors had financial ties with tobacco manufacturers. In the disputed topic of whether third-generation contraceptive pills cause an increase in thromboembolic disease, studies funded by the pharmaceutical industry find that they don't and studies funded by public money find that they do.¹⁸⁴

Effect Modification

Populations and individuals with different characteristics may be more or less susceptible to the adverse health effects of an agent. Such characteristics are called effect modifiers.

For example, sex and gender are suspected to modify the effect of a number of associations. Stronger associations between exposure to air pollution and respiratory outcomes have been reported among women relative to men.¹⁸⁵ Biologic sex may also modify associations between exposure to agents, such as pesticides and hormones, and some health outcomes.¹⁸⁶ Age may also modify associations. For example, associations between an agent and age may occur if

429 (2019), <https://doi.org/10.1007/s11606-018-4784-0>; Adam G. Dunn et al., *Conflict of Interest Disclosure in Biomedical Research: A Review of Current Practices, Biases, and the Role of Public Registries in Improving Transparency*, 1 *Rsch. Integrity & Peer Rev.* 1 (2016), <https://doi.org/10.1186/s41073-016-0006-7>.

Epidemiologists, like other scientists, also may engage in scientific misconduct, even fraud. See generally James R. Wible, *The Economics of Scientific Misconduct* 24–38 (2023) (describing two “waves” of scientific misconduct in the late 20th and early 21st centuries and reviewing studies attempting to estimate how often misconduct occurs); Steven L. George, *Research Misconduct and Data Fraud in Clinical Trials: Prevalence and Causal Factors*, 21 *Int'l J. Clinical Oncology* 15, 16 (2016), <https://doi.org/10.1007/s10147-015-0887-3> (listing several instances of falsified or fabricated data in clinical trials); Steven S. Coughlin et al., *Ethics and Scientific Integrity in Public Health, Epidemiological and Clinical Research*, 34 *Pub. Health Rev.* 1, 6 (2012), <https://doi.org/10.1007/BF03391657> (providing example of published falsified epidemiologic data); Ferric Fang et al., *Misconduct Accounts for the Majority of Retracted Scientific Publications*, 42 *PNAS* 17028, 17029 (2012) (“nearly half of retractions are for fraud or suspected fraud”); Douglas L. Weed, *Preventing Scientific Misconduct*, 88 *Am. J. Pub. Health* 125, 126 (1998), <https://doi.org/10.2105/ajph.88.1.125> (noting that although the frequency of scientific misconduct is difficult to quantify, “big effects may arise from small numbers”). Courts should be aware of this possibility, although in an adversarial system, the opposing party has the primary role in identifying such misbehavior when it could have an impact on the case at hand. In-depth discussion of scientific misconduct is beyond the scope of this guide.

184. Richard Smith, *Making Progress with Competing Interests*, 325 *BMJ* 1375, 1376 (2002), <https://doi.org/10.1136/bmj.325.7377.1375>.

185. See Jane Clougherty, *A Growing Role for Gender Analysis in Air Pollution Epidemiology*, 118 *Env't Health Persp.* 167 (2010), <https://doi.org/10.1289/ehp.0900994>.

186. See Jonathan Chevrier et al., *Sex and Poverty Modify Associations Between Maternal Peripartum Concentrations of DDT/E and Pyrethroid Metabolites and Thyroid Hormone Levels in Neonates Participating in the VHEMBE Study*, 131 *Env't Int'l* 131 (2019), <https://doi.org/10.1016/j.envint.2019.104958>.

older individuals are more likely than younger individuals to develop a particular disease (such as cancer) after exposure to the agent.

Exogenous agents may serve as effect modifiers as well. One example involves warfarin, a drug used to prevent clotting in individuals at risk of strokes and other blood-mediated adverse outcomes. A case study reported on a woman whose measure of the effect of warfarin on her blood clotting, known as the international normalized ratio (INR), was stable but whose INR became doubled when she drank cranberry juice. Her INR returned to normal when she ceased drinking cranberry juice.¹⁸⁷ Cranberry juice thus appears to be an effect modifier for the anti-coagulation effects of warfarin.

Susceptibility to a given exposure may also be modified by genetic factors, a concept known as gene–environment interaction. For example, while smoking is a known risk factor for bladder cancer, a meta-analysis¹⁸⁸ found that the relative risk was 40% greater among individuals who eliminate some chemicals more slowly owing to a particular variant of the gene NAT2, relative to those with other variants of the gene who eliminate chemicals faster.¹⁸⁹

Genetic variations¹⁹⁰ that affect how toxicants are metabolized, and therefore affect the effects of exposures on disease risks, may be common.¹⁹¹ To the

187. See Gale L. Hamann et al., *Warfarin-cranberry juice interaction*, 45(3) Ann. Pharmacotherapy e17 (2011), <https://doi.org/10.1345/aph.1P451>.

188. See section titled “Methods for Synthesizing or Combining the Results of Multiple Studies” below.

189. See Montserrat García-Closas et al., *NAT2 Slow Acetylation, GSTM1 Null Genotype, and Risk of Bladder Cancer: Results from the Spanish Bladder Cancer Study and Meta-Analyses*, 366 Lancet 649 (2005), [https://doi.org/10.1016/S0140-6736\(05\)67137-1](https://doi.org/10.1016/S0140-6736(05)67137-1).

190. A gene is a segment of DNA that, through its sequence of components called bases, codes for the production of a protein or part of a protein. Genetic variations are alterations in the sequence of bases that constitute a person’s DNA. These variations may be present in the sperm and egg cells that join to form an embryo (germline mutations), in which case they affect all of a person’s cells, or they may arise at some point during life (somatic mutations), in which case they affect only cells descended from cells in which the change occurred. DNA may vary in many different ways. A frequently studied type of genetic variation is a polymorphism—a segment of DNA that occurs in multiple forms (alleles) in the population as a result of inherited mutations in some portion of the base sequence. Polymorphisms may affect the structure or function of the protein encoded by a gene and may therefore affect health. In addition to changes in the genetic structure itself, epigenetic changes—alterations in the molecules that control the extent of gene expression (protein production)—may affect how the body interacts with potential toxicants. See, e.g., *McMillan v. Togus Regional Off., Dep’t of Veterans Affs.*, 294 F. Supp. 2d 305, 312 (E.D.N.Y. 2003) (“It is generally accepted that genetic susceptibility plays a key role in determining the adverse effects of environmental chemicals. . . . [I]f polymorphisms of the gene encoding the AhR [protein] exist in humans as they do in laboratory animals, some people would be at greater risk or at lesser risk for the toxic and carcinogenic effects of TCDD [dioxin].”)

191. See Ebony B. Bookman et al., *Gene–Environment Interplay in Common Complex Diseases: Forging an Integrative Model: Recommendations from an NIH Workshop*, 35 Genetic Epidemiology 217, 218–19 (2011), <https://doi.org/10.1002/gepi.20571> (stating that susceptibility to most human diseases “is complex and multifactorial” and describing the difficulty of assessing risk contribution in

extent that genetic variability is reflected in risk heterogeneity, an epidemiologic study that does not consider its subjects' genotypes in addition to their exposure status may produce results that are not entirely valid with respect to some portion of the study participants, even if the study validly reports the relative risk of exposure for the study sample as a whole.¹⁹² The study might understate the relative risk for people whose genetics make them more vulnerable to the effects of the exposure (perhaps even to the point of finding no association at all in the sample as a whole) and overstate the relative risk for people whose genetics protect them from the effects of the exposure. This does not mean that epidemiologic studies must consider genetic variability in order to provide useful causal evidence. Currently, most epidemiologic studies do not gather detailed genetic information on the study participants¹⁹³ for various reasons, including incomplete current scientific understanding of genetic differences in susceptibility to exposure.¹⁹⁴

gene-environment interactions); Frederica P. Perera, *Molecular Epidemiology: On the Path to Prevention?*, 92 J. Nat'l Cancer Inst. 602, 608 (2000), <https://doi.org/10.1093/jnci/92.8.602> ("new molecular epidemiologic and other data invalidate the assumption of population homogeneity"); see also, e.g., Brenda Eskenazi et al., *Organophosphate Pesticide Exposure, PON1, and Neurodevelopment in School-Age Children from the CHAMACOS Study*, 134 Env't Rsch. 149, 156 (2014), <https://doi.org/10.1016/j.envres.2014.07.001> ("Future policies regarding protection from [organophosphate] pesticide exposure should consider that some children may be more vulnerable to early life exposures by virtue of their or their mother's genetic predispositions. . . .").

192. The relative risk for the population as a whole will depend on the relative risk of exposure for each relevant genotype and the frequency of each relevant genotype in the population. Either of these values might be inaccurately measured in a study, either because of random error or because of some type of bias (e.g., a sample of people of one ancestry might not accurately reflect the genotype frequencies in a more diverse population).

193. See Witte & Thomas, *supra* note 50, at 964 ("The incorporation of environmental factors into studies of genetics has been slower moving. While some complex diseases have long been recognized as having both environmental and familial components, our understanding of their interaction is only in its infancy. . . . much remains to be understood.").

194. Even if, for instance, a study shows that the relative risk associated with an exposure varies with different variants of a particular gene, variants affecting the relative risk could also exist in other genes. The study might or might not assess what those other genes are and how differences in the various involved genes interact with each other. The biological effects of exposure to benzene, for example, appear to be influenced by variations in at least five genes. See Diana Dougherty et al., *NQO1, MPO, CYP2E1, GSTT1 and GSTM1 Polymorphisms and Biological Effects of Benzene Exposure—A Literature Review*, 182 Toxicology Letters 7, 15 (2008) ("this review shows that multiple genetic polymorphisms on the benzene metabolism pathway should be taken into account when studying the biological effects of benzene exposure"). Studies have also identified several genes with variants that are associated with different relative risks for tobacco smokers of both lung cancer, see Boffetta, *supra* note 57, at 124, and chronic obstructive pulmonary disease, see Chime-dikhamsuren Ganbold et al., *The Cumulative Effect of Gene-Gene Interactions Between GSTM1, CHRNA3, CHRNA5 and SOD3 Gene Polymorphisms Combined with Smoking on COPD Risk*, 16 Int'l J. Chronic Obstructive Pulmonary Disease 2857, 2866 (2021), <https://doi.org/10.2147/COPD>

Although some courts have considered general claims of differential susceptibility to toxic exposures,¹⁹⁵ few have been confronted with epidemiologic studies that quantitatively assess the degree to which an association between disease risk and exposure status appears to have been modified by genetic variability.¹⁹⁶ It is reasonable to believe that courts will see more such studies¹⁹⁷ offered as evidence on the issue of general causation.¹⁹⁸ If that occurs, knowledge of the plaintiff's genotype could be relevant,¹⁹⁹ although not necessarily essential, evidence on the issue of specific causation.²⁰⁰

.S320841/ ("few studies consider gene-gene or gene-environment interaction with the genetic factors in COPD susceptibility").

195. See, e.g., *In re Asbestos Prods. Liab. Litig.*, No. 10-CV-61118, 2011 WL 605801 (E.D. Pa. Feb. 16, 2011) (denying motion to exclude testimony of expert witness who opined "that susceptibility to asbestos-related disease varies within the population and the dose of asbestos necessary to cause the disease is different for each individual, and largely dependent upon genetic predisposition"); *Young v. Burton*, 567 F. Supp. 2d 121, 137 (D.D.C. 2008) (rejecting plaintiff's expert's claim that variations in HLA-DR gene affected susceptibility to illness resulting from exposure to mold); *Blackwell v. Wyeth*, 971 A.2d 235, 252–53 (Md. 2009) (affirming exclusion of expert testimony that mercury in vaccine caused autism in genetically susceptible plaintiff where trial judge found no evidence that the genetic variants about which the expert testified were associated with autism or affected mercury excretion).

196. Genetic variation in susceptibility to the toxic effects of beryllium was central to *Sheridan v. NGK Metals Corp.*, 609 F.3d 239, 244 (3d Cir. 2010), "because only persons who have a particular genetic 'marker'—the Human Leukocyte Antigen (HLA)-DPB1 allele—can potentially recognize beryllium in the lungs as an antigen." The court observed that "[m]ultiple studies have attempted to determine the percentage of the population that is genetically predisposed, or 'susceptible,' to [chronic beryllium disease]. The results so far are inconclusive and disputed." *Id.* at 245; see also *Paz v. Brush Engineered Metals, Inc.*, 555 F.3d 383, 396 n.22 (5th Cir. 2009) (quoting an expert witness's article stating that from 1% to 16% of individuals exposed to beryllium develop disease, depending on genetic susceptibility and nature of exposure).

197. See generally, e.g., Cosetta Minelli et al., *Interactive Effects of Antioxidant Genes and Air Pollution on Respiratory Function and Airway Disease: A HuGE Review*, 173 Am. J. Epidemiology 603, 618 (2011), <https://doi.org/10.1093/aje/kwq403> ("The study of genetic susceptibility can greatly improve our understanding of air pollution pathophysiologic mechanisms of action and allow identification of those pollution components with the highest potential for harm.").

198. See section titled "General Causation" below for a discussion of general causation.

199. See *Spahn v. Sec'y of Health & Hum. Servs.*, 133 Fed. Cl. 588, 593, 600 (Fed. Cl. 2017) (sustaining special master's rejection of claim for compensation for vaccine injury where claimant's expert opined that CPOX4 mutation rendered claimant ultra-susceptible to neurotoxic effects of mercury in the vaccine but claimant did not undergo genetic testing to determine whether he had the mutation).

200. Consider, hypothetically, a gene that occurs in two forms (alleles), *A* and *a*; people with at least one *A* allele (genotype *AA* or *Aa*) who are exposed to toxicant *T* have a relative risk of 10 for a particular disease, but exposed people without an *A* allele (genotype *aa*) are protected from the toxicity and have a relative risk of 1 (no association). An exposed *AA* or *Aa* plaintiff would have a different case from an exposed *aa* plaintiff. But suppose that *a* alleles are rare, so that *aa* genotypes account for only 1% of the population. It might be reasonable to permit the factfinder to draw the inference that a sick, exposed plaintiff was not *aa*, even if the plaintiff's actual

To illustrate effect modification, below are hypothetical data from a case-control study for the relationship of a dichotomous exposure, such as smoking (ever vs. never), and a dichotomous outcome, such as the existence of a disease. The data reveal effect modification by sex.

TOTAL

	Diseased	Not diseased	Total
Exposed	300	250	550
Not Exposed	200	250	450

The odds ratio for the association between exposure and disease for the total sample (males and females together) is 1.50; statistical analysis reveals a 95% confidence interval (CI) bounded by 1.20 and 1.90 and a p -value equaling 0.002 (often abbreviated as “95% CI = 1.20, 1.90; $p = 0.002$ ”). The low p -value means that this association is statistically significant for the entire study sample. However, we may have reason to believe that the exposure may differentially affect males and females. To examine this possibility, we stratify the data by sex.

MALES

	Diseased	Not diseased	Total
Exposed	80	50	130
Not Exposed	45	75	120

The odds ratio for the association between exposure and disease in the males is 2.70 (95% CI = 1.60, 4.60, $p < 0.001$). The low p -value means that this association is statistically significant, but the relatively wide confidence interval reflects the moderate size of the subsample of males ($n = 250$).

FEMALES

	Diseased	Not diseased	Total
Exposed	220	200	420
Not Exposed	155	175	330

genotype for some reason cannot be ascertained. Alternatively, limitations of the genetic epidemiology studies themselves, *see supra* note 53 and accompanying text, might affect the probative value of evidence of the plaintiff’s genotype. Specific causation is discussed in the section titled “Specific Causation” below.

The odds ratio for the association between exposure and disease in the females is 1.20 (95% CI = 0.90, 1.70, $p = 0.14$). The higher p -value means that this association is not statistically significant and the narrower confidence interval relative to males reflects the larger size of the subsample of females ($n = 750$).

We can test for whether the association between exposure and outcome is statistically different among males ($OR = 2.70$) relative to females ($OR = 1.20$). The p -value for the interaction is small (0.01), i.e., the interaction is statistically significant, suggesting that the finding that sex modifies the association between the exposure and the disease is not likely the result of random error.²⁰¹

General Causation

Once epidemiologists conclude that an association exists, the next question is whether an inference of causation is appropriate.²⁰² To make a judgment about causation, a knowledgeable expert considers the available studies and evaluates the body of evidence using the criteria described below.²⁰³

201. One court quoted an expert witness who explained the importance of effect modification thusly: “If, for example, the association between an exposure and an outcome is different for men than for women, sex modifies the relationship between the exposure and the outcome.” . . . When effect modification is present, “[c]ombining the two groups to create a summary measure of association is meaningless: it is not true for men and it is not true for women.” *Ohio Valley Envt Coal. v. Fola Coal Co.*, 120 F. Supp. 3d 509, 520–21 (S.D. W. Va. 2015) (quoting epidemiologist David Garabrandt; citations omitted).

Effect modification can coexist with confounding (see the section titled “Confounding Bias” above). Using slightly different terminology, one epidemiology textbook explained:

Effect-measure modification differs from confounding in several important ways. The most salient difference is that, whereas confounding is a bias that the investigator hopes to prevent or remove from the effect estimate, effect-measure modification is a property of the effect under study. Thus, effect-measure modification is a finding to be reported rather than a bias to be avoided or removed. Moreover, in epidemiologic analysis, one tries to eliminate confounding, but one may be interested in describing effect-measure modification.

Sebastien Haneuse & Kenneth J. Rothman, *Stratification and Standardization*, in Lash et al., *supra* note 50, at 609, 610.

202. For an excellent example of the authors of a study analyzing whether an inference of causation is appropriate in a case-control study examining whether bromocriptine (Parlodel) (a lactation suppressant) causes seizures in postpartum women, see Kenneth J. Rothman et al., *Bromocriptine and Puerperal Seizures*, 1 *Epidemiology* 232, 236–38 (1990), <https://doi.org/10.1097/00001648-199005000-00009>.

203. In a lawsuit, this would be done by an expert witness. In science, the effort is usually conducted by a panel of experts. See Douglas L. Weed, *Epidemiologic Evidence and Causal Inference*, 14 *Hematology/Oncology Clinics N. Am.* 797 (2000) (discussing judgment involved in selecting criteria and assigning rules of evidence to them in the process of judging whether an association is causal); Restatement (Third) of Torts: Liability for Physical and Emotional Harm § 28 cmt. c (2010) (“[A]n evaluation of data and scientific evidence to determine whether an inference of causation is appropriate requires judgment and interpretation.”).

When epidemiologists evaluate whether a cause–effect relationship exists between an agent and a health outcome, they are using the term causation in a way similar, but not identical, to the way that the familiar but-for, or sine qua non, test is used in law for cause in fact. “Conduct is a factual cause of [harm] when the harm would not have occurred absent the conduct.”²⁰⁴ This is equivalent to describing the act or occurrence as a necessary link in a chain of events that results in the particular outcome.²⁰⁵ Epidemiologists use causation to mean that an increase in the incidence of disease or an adverse health outcome among the group of exposed subjects would not have occurred had they not been exposed to the agent.²⁰⁶ Thus, exposure is a necessary condition for the increase in the incidence of disease among those exposed.²⁰⁷ The relationship between the epidemiologic concept of cause and the legal question of whether exposure to an agent caused an individual’s disease is addressed in the section below titled “Specific Causation.”

Data from epidemiologic studies, as well as from toxicology and in vitro studies, are also considered in an approach termed weight of evidence.²⁰⁸ In the landmark 1964 Surgeon General’s report on smoking and health, smoking was determined to be a causal agent in lung cancer based on the consistency, strength,

More generally, causation always requires an inference from circumstantial evidence because causation cannot be perceived directly with any of the five senses. Epidemiologic evidence is one form of circumstantial evidence that informs the inferential process when the causal question involves agents and disease.

204. Restatement (Third) of Torts: Liability for Physical and Emotional Harm § 26 (2010); *see also* Dan B. Dobbs et al., *Dobbs’ Law of Torts* § 189 (updated 2024). When multiple causes are each operating and each is independently capable of causing an event, the but-for, or necessary-condition, concept for causation is problematic. This is the familiar two-fires scenario, in which two independent fires simultaneously burn down a house and is sometimes referred to as overdetermined outcomes. Neither fire is a but-for, or necessary condition, for the destruction of the house, because either fire would have destroyed the house. *See* Restatement (Third) of Torts: Liability for Physical and Emotional Harm § 28 (2010). This two-fires situation is analogous to an individual being exposed to two agents, each of which is capable of causing the disease contracted by the individual. *See* *Basko v. Sterling Drug, Inc.*, 416 F.2d 417 (2d Cir. 1969).

205. *See supra* note 10; *see also* Restatement (Third) of Torts: Liability for Physical and Emotional Harm § 26 cmt. c (2010) (employing a “causal set” model to explain multiple elements, each of which is required for an outcome).

206. Bruce G. Charlton, *Attribution of Causation in Epidemiology: Chain or Mosaic?*, 49 J. Clinical Epidemiology 105, 105 (1999), [https://doi.org/10.1016/0895-4356\(95\)00030-5](https://doi.org/10.1016/0895-4356(95)00030-5) (“The imputed causal association is at the group level, and does not indicate the cause of disease in individual subjects.”).

207. *See* Lash et al., *supra* note 50, at 8 (“We can define a cause of a specific disease event as an antecedent event, condition, or characteristic that was necessary for the occurrence of the disease at the moment it occurred, given that other conditions are fixed.”); *Allen v. United States*, 588 F. Supp. 247, 405 (D. Utah 1984) (quoting a physician on the meaning of the statement that radiation causes cancer), *rev’d on other grounds*, 816 F.2d 1417 (10th Cir. 1987).

208. *See supra* note 71 and accompanying text.

specificity, temporal relationship, and coherence of the association.²⁰⁹ A year later, Sir Austin Bradford Hill published a list of elements to consider for inferring causation from an association that partially overlapped with those applied in the Surgeon General's report.²¹⁰ The items Hill identified are sometimes erroneously presented as mandatory "criteria" rather than as factors or considerations, but Hill himself acknowledged that they could only serve to assist in the inferential process: "None of my nine viewpoints can bring indisputable evidence for or against the cause-and-effect hypothesis and none can be required as a *sine qua non*."²¹¹ Yet it is imperative to note that epidemiologists employ these guidelines only after a study finds an association, to determine whether that association reflects a true causal relationship.²¹² These guidelines consist of several key inquiries that assist researchers in making a judgment about causation.²¹³ Generally, researchers are conservative when it comes to assessing causal relationships, often calling for stronger evidence and more research before a conclusion of causation is drawn.²¹⁴

209. U.S. Dep't Health, Educ. & Welfare, Pub. Health Serv., *Smoking and Health: Report of the Advisory Committee to the Surgeon General of the Public Health Service* 26–32 (1964) (describing the types of evidence considered in reaching the report's conclusion), <https://perma.cc/9TH2-JP7B>.

210. See Austin Bradford Hill, *The Environment and Disease: Association or Causation?*, 58 *Proc. Royal Soc'y Med.* 295 (1965). For discussion of these considerations and their respective strengths in informing a causal inference, see Celentano & Szklo, *supra* note 37, at 236–39; David E. Lilienfeld & Paul D. Stolley, *Foundations of Epidemiology* 276–80 (3d ed. 2019); Weed, *supra* note 203.

211. See Hill, *supra* note 210, at 36.

212. In a number of cases, experts attempted to use these guidelines to support the existence of causation in the absence of any epidemiologic studies finding an association. *Compare* *Rains v. PPG Indus.*, 361 F. Supp. 2d 829, 836–37 (S.D. Ill. 2004) (explaining Hill factors and proceeding to apply them even though there was no epidemiologic study that found an association) *and* *Wendell v. GlaxoSmithKline LLC*, 858 F.3d 1227, 1235–36 (9th Cir. 2017) (reversing exclusion of testimony of plaintiff's expert who "used the Bradford Hill methodology" but "did not rely on animal or epidemiological studies") *with* *Hoeffling v. U.S. Smokeless Tobacco Co.*, 576 F. Supp. 3d 262, 273 (E.D. Pa. 2021) (finding an expert's general causation opinion based on the Hill factor of "biological plausibility" unreliable because there was no epidemiologic evidence supporting an association), *In re Roundup Prods. Liab. Litig.*, 390 F.3d 1102, 1131 (N.D. Cal. 2018) ("identifying an association . . . is a necessary predicate to reliable application of the Bradford Hill criteria"), *and* *Jones v. Novartis Pharms. Corp.*, 235 F. Supp. 3d 1244, 1268 (N.D. Ala. 2017) (inability to point to a study showing an association between exposure and disease was "fatal flaw" of expert's reliance on Hill factors (citing the third edition of this reference guide)).

213. See Mervyn Susser, *Causal Thinking in the Health Sciences: Concepts and Strategies in Epidemiology* (1973); *Gannon v. United States*, 571 F. Supp. 2d 615, 624 (E.D. Pa. 2007) (quoting expert who testified that the Hill factors are "'well-recognized' and widely used in the science community to assess general causation"); *Chapin v. A & L Parts, Inc.*, 732 N.W.2d 578, 584 (Mich. Ct. App. 2007) (expert testified that Hill factors are the most well-utilized method for determining if an association is causal).

214. *Berry v. CSX Transp., Inc.*, 709 So. 2d 552, 568 n.12 (Fla. Dist. Ct. App. 1998) ("Almost all genres of research articles in the medical and behavioral sciences conclude their discussion with qualifying statements such as 'there is still much to be learned.' This is not, as might be assumed, an expression of ignorance, but rather an expression that all scientific fields are open-ended and can

Some regulatory agencies have modified the Hill factors in a number of ways and used them as a guide. The United States Environmental Protection Agency (EPA) has employed a system called the Causal Analysis/Diagnosis Decision Information System, or CADDIS, a weight-of-evidence approach to identify causal agents.²¹⁵ The International Agency for Research on Cancer (IARC) also relies on a weight-of-evidence approach that considers epidemiologic studies, animal experiments, and mechanistic information.²¹⁶ The IARC convenes groups of experts in various fields who are asked to reach consensus and to collectively consider the strength of the association, consistency across studies and diverse populations, dose–response relationships, biological plausibility, and temporal relationships observed in epidemiologic studies, as well as the results of animal and mechanistic studies.²¹⁷

The following is a modified list of the Hill considerations²¹⁸ that inform epidemiologists in making judgments about causation (there is no threshold number that must exist²¹⁹):

1. Temporal relationship
2. Strength of the association

progress from their present state. . . .”); *Hall v. Baxter Healthcare Corp.*, 947 F. Supp. 1387, 1446–51 (D. Or. 1996) (report of Merwyn R. Greenlick, court-appointed epidemiologist). In *Cadarian v. Merrell Dow Pharms., Inc.*, 745 F. Supp. 409 (E.D. Mich. 1989), the court refused to permit an expert to rely on a study that the authors had concluded should not be used to support an inference of causation in the absence of independent confirmatory studies. The court did not address the question whether the degree of certainty used by epidemiologists before making a conclusion of cause was consistent with the legal standard. See *DeLuca*, 911 F.2d at 957 (standard of proof for scientific community is not necessarily appropriate standard for expert opinion in civil litigation); *Wells v. Ortho Pharm. Corp.*, 788 F.2d 741, 745 (11th Cir. 1986).

215. U.S. Env’t Prot. Agency, *Causal Analysis/Diagnostic Decision Information System (CADDIS)*, <https://perma.cc/2L4R-58T8> (last visited Aug. 5, 2024); see U.S. Env’t Prot. Agency, *CADDIS Volume 1: Stressor Identification—About Causal Assessment*, “An Approach to Causal Inference,” <https://perma.cc/T52H-X84S> (“We accept multiple causation concepts and all relevant evidence and methods for turning data into evidence.”).

216. For a discussion of mechanistic data, see Eaton et al., *supra* note 60, “Toxicological Processes and Target Organ Toxicity.”

217. See generally, e.g., Int’l Agency for Rsch. on Cancer, 96 IARC Monographs on the Evaluation of Carcinogenic Risks to Humans: Alcohol Consumption and Ethyl Carbamate 7–33 (2010) (describing IARC approach).

218. This list was developed by Leon Gordis, the epidemiologist coauthor of the first three editions of this reference guide. See Celentano & Szklo, *supra* note 37, at 276–79. The original Hill considerations were: 1) consistency; 2) strength; 3) specificity; 4) dose–response; 5) temporal relationship; 6) biological plausibility; 7) coherence; and 8) experiment. See Hill, *supra* note 210. These considerations are largely congruent with the factors listed in the text, although with some wording differences and the addition of the “consideration of alternative explanations” item in the text.

219. See *Cook v. Rockwell Int’l Corp.*, 580 F. Supp. 2d 1071, 1098 (D. Colo. 2006) (“Defendants cite no authority, scientific or legal, that compliance with all, or even one, of these factors is required. . . . The scientific consensus is, in fact, to the contrary. It identifies Defendants’ list of factors as some of the nine factors or lenses that guide epidemiologists in making judgments about causation. . . . These factors are not tests for determining the reliability of any study or the causal inferences drawn from it.”).

3. Dose–response relationship
4. Replication of the findings
5. Biological plausibility
6. Consideration of alternative explanations
7. Cessation of exposure
8. Specificity of the association
9. Consistency with other relevant knowledge

There is no formula or algorithm that can be used to assess whether a causal inference is appropriate based on these guidelines.²²⁰ One or more factors may be absent even when a true causal relationship exists.²²¹ Similarly, the existence of some factors does not ensure that a causal relationship exists. Drawing causal inferences after finding an association and considering these factors requires judgment and searching analysis, based on biology, of why a factor or factors may be absent despite a causal relationship, and vice versa. Although the drawing of causal inferences is informed by scientific expertise, it is not a determination that is made by using an objective or algorithmic methodology.²²²

Temporal Relationship

A temporal, or chronological, relationship must exist for causation to exist. If an exposure causes disease, the exposure must occur before the disease develops.²²³ If the exposure occurs after the disease develops, it cannot have caused the disease. Although temporal relationship is often listed as one of many factors in assessing whether an inference of causation is justified, this aspect of a temporal relationship is a necessary factor: Without exposure before the disease, causation cannot exist.²²⁴

220. See Douglas L. Weed, *Epidemiologic Evidence and Causal Inference*, 14 *Hematology/Oncology Clinics N. Am.* 797 (2000), [https://doi.org/10.1016/s0889-8588\(05\)70312-9](https://doi.org/10.1016/s0889-8588(05)70312-9).

221. See *Cook v. Rockwell Int'l Corp.*, 580 F. Supp. 2d 1071, 1098 (D. Colo. 2006) (rejecting argument that plaintiff failed to provide sufficient evidence of causation based on failing to meet four of the Hill factors).

222. See David A. Savitz & Gregory Wellenius, *Interpreting Epidemiologic Evidence: Study Design and Analysis* 199–210 (2d ed. 2016).

223. See *McClain v. Metabolife Int'l, Inc.*, 401 F.3d 1233, 1243 (11th Cir. 2005) (“[T]he chronological relationship between exposure and effect must be biologically plausible . . . [thus], if a disease or illness in an individual preceded the established period of exposure, then it cannot be concluded that the chemical caused the disease, although it may be possible to establish that the chemical aggravated a pre-existing condition or disease.”) (quoting David L. Eaton, *Scientific Judgment and Toxic Torts: A Primer in Toxicology for Judges and Lawyers*, 12 J. L. & Pol’y 5 (2003)).

224. However, exposure during the disease initiation process may cause the disease to be more severe than it otherwise would have been without the additional dose.

With specific causation, a subject dealt with in detail in the section below titled “Specific Causation,” there may be circumstances in which a temporal relationship supports the existence of a causal relationship.

If the latency period between exposure and outcome is known,²²⁵ then exposure consistent with that information may lend credence to a causal relationship. This is particularly true when the latency period is short and competing causes are known and can be ruled out. Thus, if an individual suffers an acute respiratory response shortly after exposure to a suspected agent, and other causes of that respiratory problem are known and can be ruled out, the temporal relationship strongly supports the conclusion that a causal relationship exists.²²⁶ By contrast, exposure outside a known latency period, such as an exposure very quickly followed by development of cancer, constitutes evidence against the existence of causation.²²⁷ When latency periods are lengthy, variable, or not known, and a substantial proportion of the disease is due to unknown causes, temporal relationship provides little beyond satisfying the requirement that cause precede effect.²²⁸

225. When the latency period is known—or is known to be limited to a specific time period—as is the case with the adverse effects of some vaccines—the time frame from exposure to manifestation of disease can be critical to determining causation. Thus, epidemiologic evidence identified the swine flu vaccine as a cause of Guillain-Barre syndrome (GBS) but the peak latency period was two to three weeks after vaccination and the excess risk disappeared after 10 weeks. That evidence was critical in denying recovery to a swine flu vaccinee who developed GBS 17 weeks after vaccination. See *Robinson v. United States*, 533 F. Supp. 320, 328 (E.D. Mich. 1982).

226. For courts that have relied on temporal relationships of the sort described, see *Bonner v. ISP Technologies, Inc.*, 259 F.3d 924, 930–31 (8th Cir. 2001) (giving more credence to the expert’s opinion on causation for acute response based on temporal relationship than for chronic disease that plaintiff also developed); *Heller v. Shaw Indus., Inc.*, 167 F.3d 146 (3d Cir. 1999); *Westberry v. Gislaved Gummi AB*, 178 F.3d 257 (4th Cir. 1999); *Zuchowicz v. United States*, 140 F.3d 381 (2d Cir. 1998); *Creanga v. Jarda*, 886 A.2d 633, 641 (N.J. 2005); *Alder v. Bayer Corp., AGFA Div.*, 61 P.3d 1068, 1090 (Utah 2002) (“If a bicyclist falls and breaks his arm, causation is assumed without argument because of the temporal relationship between the accident and the injury [and, the court might have added, the absence of any plausible competing causes that might instead be responsible for the broken arm].”).

227. See *De Bazan v. Sec’y of Health & Hum. Servs.*, 539 F.3d 1347, 1352 (Fed. Cir. 2008) (plaintiff’s disease occurred too soon after vaccination for it to have been causal); *In re Phenylpropanolamine (PPA) Prods. Liab. Litig.*, 289 F. Supp. 2d 1230, 1238 (W.D. Wash. 2003) (determining that expert testimony on causation for plaintiffs whose exposure was beyond known latency period was inadmissible).

228. These distinctions provide a framework for distinguishing between cases that are largely dismissive of temporal relationships as supporting causation and others that find it of significant persuasiveness. Compare cited cases, *supra* note 226, with *Moore v. Ashland Chem. Inc.*, 151 F.3d 269, 278 (5th Cir. 1998) (giving little weight to temporal relationship in case in which there were several plausible competing causes that may have been responsible for plaintiff’s disease) and *Glastetter v. Novartis Pharms. Corp.*, 252 F.3d 986, 990 (8th Cir. 2001) (giving little weight to temporal relationship in case reports involving drug and stroke).

Strength of the Association

The relative risk is a commonly used measure of association that represents the strength of an association.²²⁹ Larger relative risks (or stronger associations using other statistical measures) are often believed to be more likely to be causal than smaller ones.²³⁰ For cigarette smoking, for example, the estimated relative risk for lung cancer is very high, about 10.²³¹ That is, the risk of lung cancer in smokers is approximately 10 times the risk in nonsmokers.

A relative risk of 10, as seen with smoking and lung cancer, is so large that, aside from causality, only very strong uncontrolled bias or large random error could account for it. Although smaller relative risks can also reflect causality, epidemiologists scrutinize such associations more closely because weaker biases (which may have been ignored or are unknown) or smaller random error could account for them.

Dose–Response Relationship

If exposure to an agent causes a disease, higher exposures would generally be expected to increase the incidence or severity of that disease.²³² Therefore, whether a dose–response relationship exists is a factor epidemiologists consider in assessing whether an association is causal.²³³ This subsection describes several possible forms

229. See section titled “Relative Risk” above.

230. See Celentano & Szklo, *supra* note 37, at 277 (“The stronger the association, the more it is that the relation is causal.”); *In re Johnson & Johnson Talcum Powder Prods. Mktg., Sales Pracs. & Prods. Litig.*, 509 F. Supp. 3d 116, 162 (D.N.J. 2020) (citing the second edition of this reference guide). The use of the strength of the association as a factor, however, does not reflect a belief that weaker causal effects occur less frequently than stronger ones. See Green, *supra* note 70, at 652–53 n.39; Tyler J. VanderWeele et al., *Causal Inference and Scientific Reasoning*, in Lash et al., *supra* note 50, at 17, 18–19 (“the stronger an association, the harder it is to explain away as artifacts of biases” but “there are many weaker associations that are generally agreed to reflect causal effects”). Indeed, the apparent strength of a given agent is dependent on the prevalence of the other necessary elements that must occur with the agent to produce the disease, rather than on some inherent characteristic of the agent itself. See Lash et al., *supra* note 50, at 87–91.

231. See Doll & Hill, *supra* note 9. The relative risk of lung cancer from smoking is a function of intensity and duration of dose (and perhaps other factors). See Karen Leffondré et al., *Modeling Smoking History: A Comparison of Different Approaches*, 156 Am. J. Epidemiology 813 (2002), <https://doi.org/10.1093/aje/kwf122>. The relative risk provided in the text is based on a specified magnitude of cigarette exposure. There is evidence that the relative risk has increased over time even though smokers in more recent studies smoked fewer cigarettes per day. Pub. Health Servs., U.S. Dep’t of Health & Hum. Servs., *The Health Consequences of Smoking: 50 Years of Progress: A Report of the Surgeon General* 158–86, 293 (2014).

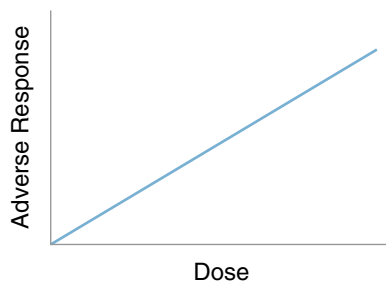
232. Exceptions to this general proposition exist. See *infra* Figure 13, depicting a non-monotonic dose–response curve, and accompanying text.

233. See *Newman v. Motorola, Inc.*, 218 F. Supp. 2d 769, 778 (D. Md. 2002) (recognizing importance of dose–response relationship in assessing causation).

of dose–response relationships. The descriptions are of adverse responses (such as the development of a disease after exposure to an agent) although some responses may be beneficial (such as the therapeutic effect of a medication).

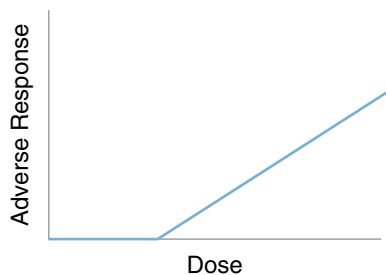
A linear dose response exists when the response increases proportionally to increasing dose, as depicted in the graph in Figure 9. The adverse response (the dependent variable) reflected in the y axis is the incidence of disease in the population and the x axis represents the dose of the agent of interest (the independent variable).²³⁴

Figure 9. Linear dose response.



Some causal agents do not exhibit a linear dose–response relationship. For example, some agents may not cause disease until the exposure exceeds a certain threshold dose. Figure 10 depicts a dose–response relationship with a threshold, with a linear dose–response relationship for doses above the threshold.

Figure 10. Linear threshold dose response.



There may be a no–effect threshold dose for some toxic substances, reflecting, for example, defense mechanisms and repair processes.²³⁵ However, for

234. Another dose–response relationship that may be of interest to researchers is the relationship of the severity of disease with increasing dose.

235. The idea that the “dose makes the poison” is a central tenet of toxicology, attributed to Paracelsus in the sixteenth century. See Eaton et al., *supra* note 60, “Introduction” section, in this

others such as carcinogens, it is commonly assumed that no threshold dose exists on a group basis.²³⁶ It is unlikely that researchers will find good evidence to support or refute the threshold–dose hypothesis because of the inability of epidemiology or animal toxicology to ascertain the effects of very small doses.²³⁷

Even if the incidence of disease increases with increasing dose, that dose–response relationship may not be linear. Supra-linear responses reflect a response that increases with increasing dose, but for which the rate of increase is smaller at higher doses than at lower doses, as depicted in Figure 11.²³⁸ Sub-linear responses reflect a response that increases with increasing dose, but for which the rate of increase is larger at higher doses than at lower doses, as depicted in Figure 12.²³⁹ Particularly for low-dose exposures, the shape of the dose–response curve—whether linear or curvilinear (and if the latter, the shape of the curve)—is a matter of hypothesis and speculation.²⁴⁰

manual. This dictum does not mean that any agent is capable of causing any disease if an individual is exposed to a sufficient dose. Rather, it reflects only the idea that there is a safe dose below which an agent does not cause any toxic effect. See Philip Wexler & Antoinette N. Hayes, *The Evolving Journey of Toxicology: A Historical Glimpse*, in Casarett and Doull's *Toxicology: The Basic Science of Poisons* 3 (Curtis D. Klaassen ed., 9th ed. 2019); see also *Alder v. Bayer Corp., AGFA Division*, 61 P.3d 1068, 1088 (Utah 2002) (illustrating misunderstanding of the concept that “the dose makes the poison”). Toxic agents tend to cause specific harmful effects, an idea that also originated with Paracelsus. See section titled “Specificity of the Association” below.

236. See, e.g., Edward J. Calabrese et al., *Ethical Challenges of the Linear Non-Threshold (LNT) Cancer Risk Assessment Revolution: History, Insights, and Lessons To Be Learned*, 2022 Sci. Total Env't 832, <https://doi.org/10.1016/j.scitotenv.2022.155054>; Bernd Kaina et al., *Do Carcinogens Have a Threshold Dose? The Pros and Cons*, in *Regulatory Toxicology* 397 (Franz-Xaver Reichl & Michael Schwenk eds., 2014). See also Irving J. Selikoff, *Disability Compensation for Asbestos-Associated Disease in the United States: Report to the U.S. Department of Labor* 181–220 (1981); *Ferebee v. Chevron Chem. Co.*, 736 F.2d 1529, 1536 (D.C. Cir. 1984) (dose–response relationship for low doses is “one of the most sharply contested questions currently being debated in the medical community”); *In re TMI Litig. Consol. Proc.*, 927 F. Supp. 834, 844–45 (M.D. Pa. 1996) (discussing low-dose extrapolation and the possibility of no-effect doses for radiation exposure and cancer).

237. Cf. Arnold L. Brown, *The Meaning of Risk Assessment*, 37 *Oncology* 302, 303 (1980); see also Advisory Comm. on Childhood Lead Poisoning Prevention, *Ctrs. for Disease Control & Prevention*, *Low Level Lead Exposure Harms Children: A Renewed Call for Primary Prevention* 7 (2012) (“new studies and re-interpretation of past studies have demonstrated that it is not possible to determine a threshold below which [blood lead level] is not inversely related to IQ”), <https://stacks.cdc.gov/view/cdc/11859>.

238. A supra-linear dose–response curve reveals a greater rate of response to increasing dose than would exist if the response rate were linear. A linear curve for the same dose and endpoint, see *supra* Figure 9, would be below the supra-linear curve depicted in Figure 11.

239. In contrast to a supra-linear curve, the responses in a sub-linear dose–response curve are lower than would exist if the response were linear. A linear curve for the same dose and endpoint, see *supra* Figure 9, would be above the sub-linear curve depicted in Figure 12.

240. See *Allen v. United States*, 588 F. Supp. 247, 419–24 (D. Utah 1984) (describing uncertainties about dose–response relationship at low levels of radiation exposure), *rev'd on other grounds*, 816 F.2d 1417 (10th Cir. 1987); *In re Bextra & Celebrex Mktg. Sales Pracs. & Prod. Liab. Litig.*, 524 F. Supp. 2d 1166, 1180 (N.D. Cal. 2007) (criticizing expert for “primitive” extrapolation of

Figure 11. Supra-linear dose response.

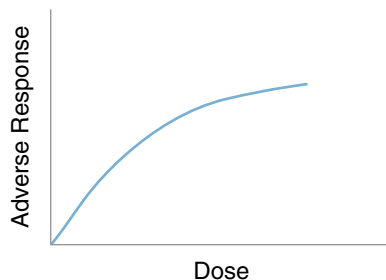
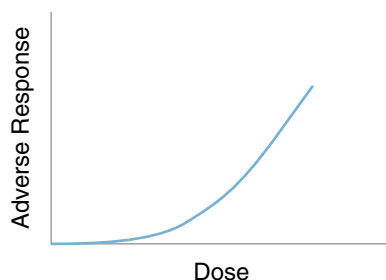


Figure 12. Sub-linear dose response.



Finally, some agents may have a greater adverse effect on health outcomes at lower doses than at higher doses. Figure 13 depicts such a non-monotonic dose-response curve.²⁴¹

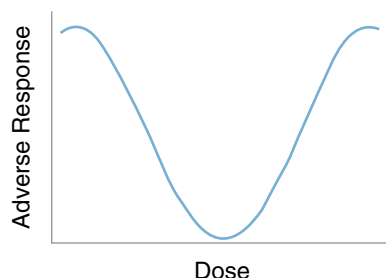
Evidence for non-monotonic dose response has been presented for chemicals that affect hormones (termed endocrine disruptors) as well as for essential nutrients. In such cases, an association may be observed at lower doses and perhaps not at moderate or high doses. For example, while there is a range of manganese intake that contributes to normal child development, low manganese intake could impair normal bone formation, reproduction, and immune response;²⁴² conversely, high exposure of pregnant women to manganese in

risk based on assumption of linear relationship of risk to dose); Troyen A. Brennan & Robert F. Carter, *Legal and Scientific Probability of Causation for Cancer and Other Environmental Disease in Individuals*, 10 J. Health Pol'y & L. 33, 43–44 (1985).

241. A non-monotonic effect is one in which the relationship between the response of the dependent variable does not continually proceed in the same direction as the independent variable, as shown in the dose-response curve in Figure 13.

242. See Longman Li & Xiaobo Yang, *The Essential Element Manganese, Oxidative Stress, and Metabolic Diseases: Links and Interactions*, *Oxidative Med. Cell Longevity*, Apr. 5, 2018, <https://doi.org/10.1155/2018/7580707>; Food & Nutrition Bd., Inst. Med., Dietary Reference Intakes for

Figure 13. Non-monotonic dose response.



drinking water may impair cognition among their children.²⁴³ Thus, a dose-response relationship is supportive, but not essential, evidence that the relationship between an agent and disease is causal.²⁴⁴

Replication of the Findings

Rarely, if ever, does a single study persuasively demonstrate a cause-effect relationship.²⁴⁵ It is important that a study be repeated in different populations and by different investigators before a causal relationship is accepted by epidemiologists and other scientists.²⁴⁶

Vitamin A, Vitamin K, Arsenic, Boron, Chromium, Copper, Iodine, Iron, Manganese, Molybdenum, Nickel, Silicon, Vanadium, and Zinc (2001); Judy L. Aschner & Michael Aschner, *Nutritional Aspects of Manganese Homeostasis*, 26 *Molecular Aspects Med.* 353 (2005), <https://doi.org/10.1016/j.mam.2005.07.003>; Pan Chen et al., *Manganese Metabolism in Humans*, 23 *Frontiers Bioscience-Landmark* 1655 (2018), <https://doi.org/10.2741/4665>.

243. See Kaitlin Vollet, *Manganese Exposure and Cognition Across the Lifespan: Contemporary Review and Argument for Biphasic Dose-Response Health Effects*, 3 *Current Env't Health Reps.* 392 (2016), <https://doi.org/10.1007/s40572-016-0108-x>.

244. Evidence of a dose-response relationship as bearing on whether an inference of general causation is justified is analytically distinct from determining whether evidence of the dose to which a plaintiff was exposed is required in order to establish specific causation. On the latter matter, see section titled “Specific Causation” below; Restatement (Third) of Torts: Liability for Physical and Emotional Harm § 28 cmt. c(2) & rptrs. note (2010).

245. In *Kehm v. Procter & Gamble Co.*, 580 F. Supp. 890, 901 (N.D. Iowa 1982), *aff'd*, 724 F.2d 613 (8th Cir. 1983), the court remarked on the persuasive power of multiple independent studies, each of which reached the same finding of an association between toxic shock syndrome and tampon use.

246. See Celentano & Szklo, *supra* note 37, at 277 (“Replication of findings is particularly important in epidemiology.”); *In re Mirena IUS Levonorgestrel-Related Prods. Liab. Litig.* (No. II), 341 F. Supp. 3d 213, 250 (S.D.N.Y. 2018) (citing the third edition of this reference guide), *aff'd*, 982 F.3d 113 (2d Cir. 2020). *But see* *Smith v. Wyeth-Ayerst Labs. Co.*, 278 F. Supp. 2d 684, 710 n.55 (W.D.N.C. 2003) (observing that replication is difficult to establish when there is only one study that has been performed at the time of trial).

The importance of replicating research findings permeates most fields of science. In epidemiology, research, if repeated, often is repeated in different populations. Consistency of the findings of multiple investigations is an important factor in making a judgment about causation. Different studies that examine the same exposure–disease relationship generally should yield qualitatively similar, but likely not quantitatively identical, results.²⁴⁷ However, it is important to keep in mind that relations between exposures and outcomes may actually differ between populations because of differences in other factors (e.g., genetics, sociodemographic characteristics, nutrition).

Biological Plausibility

Biological plausibility is not an easy criterion to use and depends upon existing knowledge about the mechanisms by which the disease develops. When biological plausibility exists, it lends credence to an inference of causality.²⁴⁸ For example, the conclusion that high cholesterol is a cause of coronary heart disease is made more plausible because cholesterol is found in atherosclerotic plaques.²⁴⁹ However, observations have been made in epidemiologic studies that were not

247. See Valentin Amrhein et al., *Inferential Statistics as Descriptive Statistics: There Is No Replication Crisis if We Don't Expect Replication*, 73 Am. Statistician 262 (2019), <https://doi.org/10.1080/00031305.2018.1543137>.

248. A number of courts have adverted to this criterion in the course of their discussions of causation in toxic substances cases. *E.g.*, *In re Phenylpropanolamine (PPA) Prods. Liab. Litig.*, 289 F. Supp. 2d 1230, 1247–48 (W.D. Wash. 2003); *Cook v. United States*, 545 F. Supp. 306, 314–15 (N.D. Cal. 1982) (discussing biological implausibility of a two-peak increase of disease when plotted against time); *Landrigan v. Celotex Corp.*, 605 A.2d 1079, 1085–86 (N.J. 1992) (discussing the existence vel non of biological plausibility); see also Eaton et al., *supra* note 60, “Demonstrating an Association Between Exposure and Risk of Disease.”

249. High LDL cholesterol levels are associated with higher rates of coronary heart disease. The physical presence of cholesterol in plaques found in blood vessels lends credence—biological plausibility—to the idea that this association is causal. The causal relation has been confirmed by randomized clinical trials of statin drugs, which reduce blood levels of LDL cholesterol and have a protective effect against heart disease.

Epidemiologists have also used another approach to help rule out the possibility that the association resulted from the effects of some unknown confounder (a factor causing both high LDL cholesterol and coronary heart disease) or even the possibility that coronary heart disease causes LDL cholesterol to collect in plaques (i.e., so-called reverse causation). This approach relies on genetic variability to provide information about the causal role of cholesterol levels.

Some people have genetic variations that affect their metabolism of fats and reduce their LDL cholesterol levels. People with these cholesterol-lowering genes have reduced risk of coronary heart disease. The observation that an inherited genetic variation is associated with both lower LDL cholesterol levels and reduced risk of coronary heart disease strongly supports the inference that cholesterol levels affect the disease risk rather than heart disease causing cholesterol or confounding. Because the presence or absence of coronary heart disease does not affect an individual's inherited genes, there is no possibility of reverse causation. Because some unknown factor does not alter both

biologically plausible at the time but that were subsequently shown to be correct.²⁵⁰ When an observation is inconsistent with current biological knowledge, it should not be discarded, but the observation should be confirmed before it is given credence. The salience of this factor varies depending on the extent of scientific knowledge about the toxicology as well as cellular and subcellular mechanisms through which the disease process works. The mechanisms of some diseases are understood quite well based on evidence, including from toxicologic research, whereas other mechanism explanations are merely hypothesized—although hypotheses are sometimes accepted under this factor.²⁵¹

a person's inherited genes and a person's coronary heart disease risk, the possibility of confounding is greatly reduced (see section titled "Effect Modification" above).

This epidemiologic method—known as Mendelian randomization (named after Gregor Mendel, the pioneering researcher of the laws of heredity)—uses genetic variation (in the example, the cholesterol-lowering variant) as a proxy for an exposure of interest (in the example, LDL cholesterol levels), to help clarify whether the exposure is causally associated with a health outcome of interest (in the example, coronary heart disease). As vast genomic datasets are increasingly available, the number of studies applying Mendelian randomization is increasing rapidly.

Mendelian randomization can be applied when genetic variations are associated with the exposure of interest and certain other assumptions are satisfied. It might seem implausible that a genetic variation is associated with the likelihood that someone is exposed to a potentially toxic substance. But genetic variation could affect how a person metabolizes a substance, which in turn could affect the substance's toxicity in that person. In such a situation, Mendelian randomization could be useful for the causal issues discussed in this reference guide; we thus provide this brief introduction to the methodology. As of the date of publication of this reference guide, however, Mendelian randomization has not been prominent in research on the types of causal issues addressed here and has not figured at all in court opinions concerning those issues. For short introductions to the methodology, see Gemma Sharp, *Introduction to Mendelian Randomization*, youtube.com/watch?v=isMZhasFc4M (Mar. 10, 2020) (last visited November 13, 2024); Connor A. Emdin et al., *Mendelian Randomization*, 318 JAMA 1925 (2017), <https://doi.org/10.1001/jama.2017.17219>. For a more mathematical introduction, see E. Sanderson et al., *Mendelian Randomization*, 2 Nature Rev. Methods Primers [6] (2022), <https://doi.org/10.1038/s43586-021-00092-5>.

250. For example, Sir Norman Gregg connected German measles in pregnant women with congenital cataracts before there was understanding of the role of viruses as teratogens. See Celenzano & Szklo, *supra* note 37, at 277; N. McAllister Gregg, *Congenital Cataract Following German Measles in the Mother*, 107 Epidemiology & Infection iii (1991), <https://doi.org/10.1017/s0950268800048627>.

251. See Douglas L. Weed & Stephen D. Hursting, *Biologic Plausibility in Causal Inference: Current Methods and Practice*, 147 Am. J. Epidemiology 415 (1998), <https://doi.org/10.1093/oxfordjournals.aje.a009466> (examining use of this criterion in contemporary epidemiologic research and distinguishing between alternative explanations of what constitutes biological plausibility, ranging from mere hypotheses to "sufficient evidence to show how the factor influences a known disease mechanism"); David A. Savitz, *Epidemiology and Biological Plausibility in Assessing Causality*, 5 Env't Epidemiology e177 (2021), <https://doi.org/10.1097/EE9.0000000000000177>. For a discussion of the role of mechanistic research in toxicology, see Eaton et al., *supra* note 60, "Toxicological Processes and Target Organ Toxicity."

Consideration of Alternative Explanations

This consideration refers to the possibility that an observed association is caused by something other than the exposure under study. The importance of considering the possibility of bias and confounding and ruling out those possibilities is discussed above.²⁵²

Cessation of Exposure

If an agent is a cause of a disease, then one would expect that cessation of exposure to that agent ordinarily would reduce the risk of the disease.²⁵³ This has been the case, for example, with cigarette smoking and lung cancer. In many situations, however, relevant data are simply not available regarding the possible effects of ending the exposure. But when such data are available, and eliminating exposure reduces the incidence of disease, this factor strongly supports a causal relationship.

Specificity of the Association

An association exhibits specificity if the exposure is associated only with a single disease or type of disease. This criterion reflects the fact that although an agent causes one disease, it does not necessarily cause other diseases.²⁵⁴ This

252. See sections titled “Biases” and “Effect Modification” above.

253. In his speech, Hill labeled this consideration “Experiment” because, as compared to other types of observational studies, the ability to observe what happens when an exposure ceases is more closely akin to a controlled experiment.

254. See, e.g., *Nelson v. Am. Sterilizer Co.*, 566 N.W.2d 671, 676–77 (Mich. Ct. App. 1997) (affirming dismissal of plaintiff’s claims that chemical exposure caused her liver disorder, but recognizing that evidence supported claims for neuropathy and other illnesses); *Sanderson v. Int’l Flavors & Fragrances, Inc.*, 950 F. Supp. 981, 996–98 (C.D. Cal. 1996); see also *Taylor v. Airco, Inc.*, 494 F. Supp. 2d 21, 27 (D. Mass. 2007) (holding that plaintiff’s expert could testify to causal relationship between vinyl chloride and one type of liver cancer for which there was only modest support given strong causal evidence for vinyl chloride and another type of liver cancer).

When a party claims that evidence of a causal relationship between an agent and one disease is relevant to whether the agent caused another disease, courts have required the party to show that the mechanisms involved in development of the disease are similar. For example, in *Milward v. Acuity Specialty Products*, 639 F.3d 11, 19–20 (1st Cir. 2011), the court accepted the plaintiff’s expert’s testimony that benzene caused a rare form of leukemia based on studies supporting the proposition that benzene was a cause of the general form of leukemia and testimony about the common causal mechanism for all types of leukemia. *Accord In re Bextra & Celebrex Mktg. Sales Pracs. & Prod. Liab. Litig.*, 524 F. Supp. 2d 1166, 1183 (N.D. Cal. 2007); *Magistrini v. One-Hour Martinizing Dry Cleaning*, 180 F. Supp. 2d 584, 603 (D.N.J. 2002).

consideration is highly debatable. Evidence amassed since the publication of Hill's paper runs contrary to this. For example, cigarette manufacturers have long claimed that because cigarettes have been linked to multiple adverse health outcomes including lung cancer, emphysema, bladder cancer, heart disease, pancreatic cancer, and other conditions, there is no specificity and the relationships are not causal. There is, however, at least one good reason why inferences about the health consequences of tobacco do not require specificity: Because tobacco and cigarette smoke are not in fact single agents but consist of numerous harmful agents, smoking represents exposure to multiple agents, with multiple potential effects. However, even a single chemical can increase risk for multiple disease endpoints, e.g., multiple cancers.²⁵⁵ For example, asbestos causes both lung cancer and mesothelioma.

Consistency with Other Relevant Knowledge

In addressing the causal relationship of lung cancer to cigarette smoking, researchers examined trends over time for lung cancer and for cigarette sales in the United States. A marked increase in lung-cancer death rates in men was observed with a time lag reflecting the latency period for lung cancer, which appeared to follow the increase in sales of cigarettes. Had the increase in lung-cancer deaths followed a decrease in cigarette sales, it might have given researchers pause. It would not have precluded a causal inference, but the inconsistency of the trends in cigarette sales and lung-cancer mortality would have had to be explained.

Background Rate of Disease

The background rate of disease is neither a Hill/Gordis factor²⁵⁶ nor a consideration involved in determining whether an association should be judged as causal. Still, some courts have stated that, independent of epidemiologic studies, the background rate of the disease at issue must be identified and known by an

255. See, e.g., Dario Consonni et al., *Mortality and Cancer Incidence in a Population Exposed to TCDD After the Seveso, Italy, Accident (1976-2013)*, 81 *Occup. Env't Med.* 349 (2024) (describing association of exposure with various cancers and other illnesses after an accidental release of dioxin, depending on geographic location relative to the accident, gender, and time after the accident); Marcella Warner et al., *Dioxin Exposure and Cancer Risk in the Seveso Women's Health Study*, 119 *Env't Health Persps.* 1700, 1703 (2011) (reporting "a statistically significant, dose-related increased risk in overall cancer incidence associated with" levels of dioxin in blood).

256. This refers to the considerations for assessing the causal significance of an association first described by Sir Austin Bradford Hill as modified by Leon Gordis. See *supra* note 218.

expert testifying to general causation.²⁵⁷ This is incorrect and fails to appreciate epidemiologic methodology. Epidemiologists examine the difference in rates of disease between the exposed and unexposed groups in a study.²⁵⁸ In a well-designed epidemiologic study, the “background” rate—the likelihood of disease even absent exposure—will be reflected in the rate of disease in the unexposed group.²⁵⁹ The magnitude of the background rate is unimportant for determining if that rate is different from the rate in the exposed group. To put the point slightly differently, whether the background rate is very low, moderate, or relatively high does not affect the researcher’s assessment of the data’s probative value for determining general causation.²⁶⁰

Methods for Synthesizing or Combining the Results of Multiple Studies

The scientific record often includes a number of epidemiologic studies examining a similar exposure and outcome. Often those studies will have different

257. See *In re Deepwater Horizon Belo Cases*, No. 3:19CV963-MCR-HTC, 2022 WL 17721595, at *5 (N.D. Fla. Dec. 15, 2022) (“Thus, a failure to identify or describe the background risk of a disease is a ‘serious methodological deficiency’ and ‘substantial weakness’ in an expert’s general causation opinion.”) (quoting *Chapman v. Procter & Gamble Distrib., LLC*, 766 F.3d 1296, 1307 (11th Cir. 2014)); *In re Abilify (Aripiprazole) Prods. Liab. Litig.*, 299 F. Supp. 3d 1291, 1308 (N.D. Fla. 2018); *Jones v. Novartis Pharms. Corp.*, 235 F. Supp. 3d 1244, 1280 (N.D. Ala. 2017) (“A failure to assess or identify the background risk of a disease is a ‘serious methodological deficiency’ because, ‘[w]ithout a baseline, any incidence may be coincidence.’”) (quoting *Chapman v. Procter & Gamble Distrib., LLC*, 766 F.3d 1296, 1307–08 (11th Cir. 2014)), *aff’d in part*, 720 F. App’x 1006 (11th Cir. 2018).

258. See section titled “Relative Risk” above. Alternatively, in a case-control design, epidemiologists examine the difference in the odds of exposure among groups with and without the disease. See section titled “Odds Ratio” above.

259. This is not to say that the background rate of disease is entirely unimportant to epidemiologists. If a disease is very rare, it may be difficult to conduct a study with enough subjects to generate statistically meaningful results, even if an exposure actually causes the disease. The background rate of disease also may play an important role in public health decision-making, as it can address the relative benefit of an intervention that reduces that rate of disease.

260. Some courts have implied that the presence of a relatively high background rate of a disease itself undermines an inference of general causation. To see why this is incorrect, imagine a disease that afflicts as many as 20% of the general population, but that—as revealed by a well-designed epidemiologic study—afflicts 100% of people exposed to a particular agent. In this example, the relative risk would be 5.0. We likely would have no trouble making the inference that the agent causes at least some cases of disease among those exposed, assuming that the relevant considerations epidemiologists take into account, see *supra* text accompanying notes 218–54, support that inference. The same would be equally true with an extremely low background rate of disease that was five times greater in the exposed group.

results; even with well-conducted studies, we should expect to find variation in the results.²⁶¹ Some may find an association while others may not, or studies may report associations, but of different magnitude.²⁶² Different studies may also be conducted in populations with different characteristics, apply different statistical techniques, measure exposure or the outcome differently, or control for different confounders, among other things, which may explain the different results. There are two main types of data-synthesis methodologies available when multiple studies address an issue of scientific interest: 1) systematic review²⁶³ and 2) meta-analysis.²⁶⁴ The two are similar in that they do not directly investigate the question of interest by gathering and analyzing primary data, but instead synthesize studies that have been previously performed. Systematic reviews²⁶⁵ are not quantitative; meta-analyses provide a quantitative summary of data gathered by multiple prior studies. Both are important because science is a cumulative process and all relevant prior work should be considered when addressing a scientific issue. Systematic reviews have been described as “a review of evidence relevant to a clearly formulated question that uses systematic and explicit methods to identify, select and critically appraise relevant research, and to collect and analyze data from the studies that are included within the review.”²⁶⁶

In view of the fact that studies may disagree and that studies may be based on a small number of individuals and lack the statistical power needed for firmer conclusions, the technique of meta-analysis was developed.²⁶⁷ Meta-analysis is a

261. See Valentin Amrhein et al., *Inferential Statistics as Descriptive Statistics: There Is No Replication Crisis if We Don't Expect Replication*, 73 *Am. Statistician* 262 (2019), <https://doi.org/10.1080/00031305.2018.1543137>.

262. See, e.g., *Zandi v. Wyeth a/k/a Wyeth, Inc.*, No. 27-CV-06-6744, 2007 WL 3224242 (Minn. Dist. Ct. Oct. 15, 2007) (plaintiff's expert cited 40 studies in support of a causal relationship between hormone therapy and breast cancer; many studies found different magnitudes of increased risk).

263. See generally *Cochrane Handbook for Systematic Reviews of Interventions* (Julian P.T. Higgins et al. eds., 2d ed. 2019). A set of guidelines, Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA), is used for systematic reviews to provide consistency and transparency and to improve the reliability of the reviews. See Matthew J. Page et al., *The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews*, 372 *Brit. Med. J.* n.71 (2021), <https://doi.org/10.1136/bmj.n71>.

264. See Savitz & Wellenius, *supra* note 222, at 185–98.

265. See generally Guy Paré & Spyros Kitsiou, *Methods for Literature Reviews*, in *Handbook of eHealth Evaluation: An Evidence-Based Approach* (F. Lau & C. Kuziemsky eds., 2017), <https://perma.cc/RBF7-7FXE>.

266. Env't Evidence, *Systematic Review*, Collaboration for Env't Evidence, <https://perma.cc/9RHX-KB2K>.

267. See *In re Paoli R.R. Yard PCB Litig.*, 916 F.2d 829, 856 (3d Cir. 1990); *In re Viagra (Sildenafil Citrate) & Cialis (Tadalafil) Prods. Liab. Litig.*, 424 F. Supp. 3d 781, 787 (N.D. Cal. 2020) (“Meta-analysis has the advantage of pooling more data so that the results are less likely to be misleading solely due to chance.”). Thus, contrary to the suggestion by at least one court, multiple studies with small numbers of subjects may be pooled to reduce the possibility of sampling error. See *In re Joint E. & S. Dist. Asbestos Litig.*, 827 F. Supp. 1014, 1042 (S.D.N.Y. 1993) (“[N]o matter

method of pooling study results to arrive at a single measure of association to represent the totality of the studies reviewed.²⁶⁸ It is a way of systematizing the time-honored approach of reviewing the literature, which is characteristic of science, and placing it in a standardized framework with quantitative methods for estimating associations. In a meta-analysis, studies are given different weights in proportion to the sizes of their study populations and other characteristics.²⁶⁹

Meta-analysis is most useful when employed in pooling randomized experimental trials, because in that circumstance the studies included in the meta-analysis share an important methodologic characteristic, namely the randomized assignment of subjects to different exposure groups.²⁷⁰ Meta-analysis applied to observational studies is also useful but is more challenging.²⁷¹ Studies may vary in the methods employed (e.g., research design and statistical approaches) as well as in the extent to which they are affected by and may have adjusted or corrected for confounding, selection bias, and measurement error. Such differences raise questions about the appropriateness of combining studies' results. The production of a single estimate of association, sometimes with a narrow confidence interval, may thus provide a false sense of security. Meta-analyses should be evaluated as carefully as one would evaluate the overall literature to draw a conclusion about the likelihood that an exposure causes a disease.²⁷² For

how many studies yield a positive but statistically insignificant SMR for colorectal cancer, the results remain statistically insignificant. Just as adding a series of zeros together yields yet another zero as the product, adding a series of positive but statistically insignificant SMRs together does not produce a statistically significant pattern.”), *rev'd*, 52 F.3d 1124 (2d Cir. 1995).

268. For a nontechnical explanation of meta-analysis, along with case studies of a variety of scientific areas in which it has been employed, see Morton Hunt, *How Science Takes Stock: The Story of Meta-Analysis* (1997).

269. Petitti, *supra* note 121.

270. On rare occasions, meta-analyses of both clinical and observational studies are available. See, e.g., *In re Bextra & Celebrex Mktg. Sales Pracs. & Prod. Liab. Litig.*, 524 F. Supp. 2d 1166, 1175 (N.D. Cal. 2007) (referring to clinical and observational meta-analyses of a low dose of a drug; both analyses failed to find any effect).

271. See Jesse A. Berlin et al., *The Use of Meta-Analysis in Pharmacoepidemiology*, in *Pharmacoepidemiology*, at 897, 899–901 (Brian L. Strom et al. eds., 6th ed. 2020); Donna F. Stroup et al., *Meta-Analysis of Observational Studies in Epidemiology: A Proposal for Reporting*, 283 JAMA 2008, 2009 (2000), <https://doi.org/10.1001/jama.283.15.2008>; Zachary B. Gerbarg & Ralph I. Horwitz, *Resolving Conflicting Clinical Trials: Guidelines for Meta-Analysis*, 41 J. Clinical Epidemiology 503 (1988), [https://doi.org/10.1016/0895-4356\(88\)90053-4](https://doi.org/10.1016/0895-4356(88)90053-4). See also *Deutsch v. Novartis Pharm. Corp.*, 768 F. Supp. 2d 420, 451–52 (E.D.N.Y. 2011) (rejecting defendants' claim that meta-analysis is sufficiently reliable to support an expert opinion only if the meta-analysis is applied to clinical trials).

272. Courts have recognized that a properly performed meta-analysis can provide a basis for admissible expert testimony. In *Cooper v. Takeda Pharms. Am., Inc.*, 191 Cal. Rptr. 3d 67, 95 (Ct. App. 2015), the court reinstated a jury verdict for the plaintiffs, observing that in striking plaintiffs' experts' testimony after trial, “the trial court's piecemeal rejection of individual studies

instance, it is essential to ensure that the meta-analysis is well conducted. A well-conducted meta-analysis appropriately considers 1) the quality of the included studies, 2) whether publication bias (the greater probability of publication of significant findings) is likely to be present, 3) if unpublished studies should be sought and included, 4) whether criteria for including and excluding studies are appropriate, 5) whether appropriate subgroup analyses were conducted, and 6) whether results are exceedingly heterogeneous, and if so, whether the sources of heterogeneity (such as whether the effect of an exposure is truly different in populations with different characteristics) have been appropriately investigated.²⁷³

Specific Causation

Epidemiology is concerned with the incidence of disease in populations, and epidemiologic studies do not address the question of the cause of an individual's

was inappropriate and ignored the testimony by [plaintiffs' experts] that the results of the individual studies considered as a whole, including in the meta-analyses, was what really persuaded them that Actos causes bladder cancer." Similarly reflecting the value of meta-analysis, in *In re Bextra & Celebrex Mktg. Sales Pracs. & Prod. Liab. Litig.*, 524 F. Supp. 2d 1166 (N.D. Cal. 2007), the court relied on several meta-analyses of Celebrex at a 200-mg dose to conclude that the plaintiffs' experts who proposed to testify to toxicity at that dosage failed to meet the requirements of *Daubert*. The court criticized those experts for the wholesale rejection of meta-analyses of observational studies. And, in *In re Paoli Railroad Yard PCB Litig.*, 916 F.2d 829, 856–57 (3d Cir. 1990), the court overturned the district court's exclusion of a report that used meta-analysis, observing that meta-analysis is a regularly used scientific technique. But the Third Circuit recognized that the technique might be poorly performed and required the district court to reconsider the validity of the expert's work in performing the meta-analysis. See also *In re Zolof (Sertraline Hydrochloride) Prod. Liab. Litig.*, 858 F.3d 787, 796 (3d Cir. 2017) (affirming exclusion of expert testimony that involved meta-analyses because, even assuming that meta-analysis is a reliable technique, the expert did not "reliably apply the 'technique' to the body of evidence . . ."); *Carl v. Johnson & Johnson*, 237 A.3d 308 (N.J. Super. Ct. App. Div. 2020) (reversing exclusion of plaintiffs' experts' testimony and holding that the experts, who relied heavily on meta-analyses to support their opinions, "adhered to methodologies generally followed by experts in the field and relied upon studies and information generally considered an acceptable basis for inclusion in the formulation of expert opinions").

273. See Mathew J. Page et al., *The PRISMA 2020 Statement: an Updated Guideline for Reporting Systematic Reviews*, 134 J. Clinical Epidemiology 178, 186 (2021), <https://doi.org/10.1136/bmj.n71> (consensus statement describing update to a checklist of items to be included for "more transparent, complete, and accurate reporting of systematic reviews, thus facilitating evidence based decision making"); Stroup et al., *supra* note 271, at 2009 (consensus statement providing checklist of items to report in a meta-analysis based on a systematic review of published meta-analyses); Monika Mueller et al., *Methods to Systematically Review and Meta-Analyse Observational Studies: A Systematic Scoping Review of Recommendations*, 18 BMC Med. Rsch. Methodology 44 (2018), <https://doi.org/10.1186/s12874-018-0495-9> (summarizing recommendations for conducting meta-analyses of observational studies).

disease.²⁷⁴ This question, often referred to as specific causation, is beyond the domain of the science of epidemiology.²⁷⁵

Epidemiology has its limits at the point where an inference is made that the relationship between an agent and a disease is causal (general causation) and where the magnitude of excess risk attributed to the agent has been determined; that is, epidemiologists investigate whether an agent can cause a disease, not whether an agent did cause a specific plaintiff's disease.²⁷⁶

Nevertheless, the specific causation issue is a necessary legal element in a toxic tort case. The plaintiff must establish not only that the defendant's agent is capable of causing disease generally, but also that it caused the plaintiff's disease specifically.²⁷⁷ Thus, numerous cases have confronted the question of what is acceptable proof of specific causation and the role that epidemiologic evidence plays in answering that question.²⁷⁸ But this question is not addressed by

274. See *DeLuca v. Merrell Dow Pharms., Inc.*, 911 F.2d 941, 945 & n.6 (3d Cir. 1990) ("Epidemiological studies do not provide direct evidence that a particular plaintiff was injured by exposure to a substance."); *Rhyne v. U.S. Steel Corp.*, 474 F. Supp. 3d 733, 750 (W.D.N.C. 2020) ("epidemiology focuses on the issue of general causation, not specific causation"); *In re E. I. du Pont de Nemours & Co. C-8 Pers. Inj. Litig.*, No. 2:18-CV-00136, 2019 WL 6894069, at *6 (S.D. Ohio Dec. 18, 2019) (explaining that epidemiology focuses on general causation as opposed to specific causation (citing the third edition of this reference guide)); Michael Dore, *A Commentary on the Use of Epidemiological Evidence in Demonstrating Cause-in-Fact*, 7 Harv. Env't L. Rev. 429, 436 (1983); Khristine L. Hall & Ellen K. Silbergeld, *Reappraising Epidemiology: A Response to Mr. Dore*, 7 Harv. Env't L. Rev. 441, 445 (1983).

There are, however, some diseases that do not occur without exposure to a given toxic agent. This is the same as saying that the toxic agent is a necessary cause for the disease, and the disease is sometimes referred to as a signature disease (also, the agent is pathognomonic) because the existence of the disease necessarily implies the causal role of the agent. Two examples are asbestosis, which is a signature disease for asbestos, and vaginal adenocarcinoma (in young adult women), which is a signature disease for in utero DES exposure. See Kenneth S. Abraham & Richard A. Merrill, *Scientific Uncertainty in the Courts*, in *Issues Sci. & Tech.* 93, 101 (1986).

275. See *Jones v. Novartis Pharms. Corp.*, 235 F. Supp. 3d 1244, 1252 (N.D. Ala. 2017) ("No scientific methodology exists for assessing specific causation for an individual based on group studies."), *aff'd in part*, *Jones v. Novartis Pharms. Co.*, 720 F. App'x 1006 (11th Cir. 2018).

276. See *In re E. I. du Pont de Nemours & Co. C-8 Pers. Inj. Litig.*, No. 2:18-CV-00136, 2019 WL 6894069, at *6 (S.D. Ohio Dec. 18, 2019) (quoting the third edition of this reference guide).

277. See, e.g., *Adkisson v. Jacobs Eng'g Grp., Inc.*, 342 F. Supp. 3d 791, 798 (E.D. Tenn. 2018) (explaining the difference between general causation and specific causation and observing the distinction is "well-accepted in federal courts"). Specific-causation evidence only matters when general causation has been established. See *Williams v. Mosaic Fertilizer*, No. 8:14-CV-1748-T-35MAP, 2016 WL 7175657, at *11 (M.D. Fla. June 24, 2016) ("Without an adequate basis for establishing general causation, Dr. Mink's specific causation opinions are irrelevant."), *aff'd*, 889 F.3d 1239 (11th Cir. 2018).

278. In many instances, causation can be established without epidemiologic evidence. When the mechanism of causation is well understood, the causal relationship is well established, or the timing between cause and effect is close, scientific evidence of causation may not be required. This is frequently the situation when the plaintiff suffers traumatic injury rather than disease. This

epidemiology.²⁷⁹ It is instead a legal question—one with which numerous courts have grappled.²⁸⁰ The remainder of this section is predominantly an explanation of judicial opinions and the reasoning courts have used in applying risk estimates from epidemiology to address the issue of specific causation in individual cases.

Before proceeding, one last caveat is in order. This section assumes that epidemiologic evidence has been used as proof of causation for a given plaintiff. The discussion does not address whether a plaintiff must or should use epidemiologic evidence to prove causation.²⁸¹

section addresses only those situations in which causation is not evident, and scientific evidence is required.

279. Nevertheless, an epidemiologist may be helpful to the factfinder answering this question. Some courts have permitted epidemiologists, or those who use epidemiologic methods, to testify about specific causation. See *Ambrosini v. Labarraque*, 101 F.3d 129, 137–41 (D.C. Cir. 1996); *Zuchowicz v. United States*, 870 F. Supp. 15 (D. Conn. 1994); *Landrigan v. Celotex Corp.*, 605 A.2d 1079, 1088–89 (N.J. 1992); *Carl v. Johnson & Johnson*, 237 A.3d 308, 338 (N.J. Super. App. Div. 2020). But see *Rhyne v. U.S. Steel Corp.*, 474 F. Supp. 3d 733, 750 (W.D.N.C. 2020) (concluding epidemiologist was not qualified to testify to specific causation).

In general, courts seem more concerned with the basis of an expert's opinion than with whether the expert is an epidemiologist or clinical physician. See *Porter v. Whitehall*, 9 F.3d 607, 614 (7th Cir. 1992) (“curb side” opinion from clinician not admissible); *Burton v. R.J. Reynolds Tobacco Co.*, 181 F. Supp. 2d 1256, 1266–67 (D. Kan. 2002) (vascular surgeon permitted to testify to general causation over objection based on fact he was not an epidemiologist); *Wade-Greaux v. Whitehall Labs.*, 874 F. Supp. 1441, 1469–72 (D.V.I.) (clinician's multiple bases for opinion inadequate to support causation opinion), *aff'd*, 46 F.3d 1120 (3d Cir. 1994); *Landrigan*, 605 A.2d at 1083–89 (permitting both clinicians and epidemiologists to testify to specific causation provided the methodology used is sound); *Trach v. Fellin*, 817 A.2d 1102, 1118–19 (Pa. Super. Ct. 2003) (toxicologist and pathologist permitted to testify to specific causation).

280. See Restatement (Third) of Torts: Liability for Physical and Emotional Harm § 28 cmt. c(3) (2010) (“Scientists who conduct group studies do not examine specific causation in their research. No scientific methodology exists for assessing specific causation for an individual based on group studies. Nevertheless, courts have reasoned from the preponderance-of-the-evidence standard to determine the sufficiency of scientific evidence on specific causation when group-based studies are involved. . .”).

281. See *id.* § 28 cmt. c(3) & rptrs. note (“most courts have appropriately declined to impose a threshold requirement that a plaintiff always must prove causation with epidemiologic evidence”); see also *Westberry v. Gislaved Gummi AB*, 178 F.2d 257 (4th Cir. 1999) (holding that expert's testimony on causation was properly admitted where there was an acute response, a differential diagnosis that ruled out other known causes of disease, and the absence of epidemiologic or toxicologic studies and there were dechallenge/rechallenge tests conducted by expert that were consistent with exposure to defendant's agent causing disease); *Zuchowicz v. United States*, 140 F.3d 381 (2d Cir. 1998); *In re Testosterone Replacement Therapy Prods. Liab. Litig.* Coordinated Pretrial Proc., No. 14 C 1748, 2017 WL 1833173, at *15 (N.D. Ill. May 8, 2017) (“Where the existing epidemiological research is limited or nonexistent, courts are more willing to allow expert testimony on causation that does not depend on epidemiological sources.”) (citing cases); *In re Tylenol (Acetaminophen) Mktg., Sales Prac. & Prods. Liab. Litig.*, No. 2:12-CV-07263, 2016 WL 3997046, at *7 (E.D. Pa. July 26, 2016) (rejecting defendant's contention that because there were no epidemiologic studies available, plaintiff's expert could not testify to causation of liver failure due to Tylenol use and explaining the reasons why such studies were unavailable for such a relationship).

Two legal issues arise with regard to the role of epidemiology in proving individual causation: admissibility and sufficiency of evidence to meet the burden of production. The first issue tends to receive less attention by the courts, because epidemiologic studies are rarely introduced independently into evidence,²⁸² but nevertheless deserves mention: An epidemiologic study that is sufficiently rigorous to justify a conclusion that it is scientifically valid should be admissible,²⁸³ as it tends to make an issue in dispute more or less likely.²⁸⁴

Far more courts have confronted the role that epidemiology plays with regard to the sufficiency of the evidence and the burden of production.²⁸⁵ The civil burden of proof is described most often as requiring the factfinder to “believe that what is sought to be proved . . . is more likely true than not true.”²⁸⁶ The relative risk from epidemiologic studies can be adapted to this 50%-plus standard to yield a probability or likelihood that an agent caused an individual’s disease.²⁸⁷ But while the discussion below speaks in terms of the magnitude of

282. Most often, epidemiologic and other scientific studies are used by expert witnesses as the basis of their causal opinions and are not independently admitted into evidence.

283. See *DeLuca v. Merrell Dow Pharms., Inc.*, 911 F.2d 941, 958 (3d Cir. 1990); cf. *Kehm v. Procter & Gamble Co.*, 580 F. Supp. 890, 902 (N.D. Iowa 1982) (“These [epidemiologic] studies were highly probative on the issue of causation—they all concluded that an association between tampon use and menstrually related TSS [toxic shock syndrome] cases exists.”), *aff’d*, 724 F.2d 613 (8th Cir. 1984). In *Ellis v. International Playtex, Inc.*, 745 F.2d 292, 303 (4th Cir. 1984), the court concluded that certain epidemiologic studies were admissible under an exception to the hearsay rule despite criticism of the methodology used in the studies. The court held that the claims of bias went to the studies’ weight rather than their admissibility.

284. Even if evidence is relevant, it may be excluded if its probative value is substantially outweighed by prejudice, confusion, or inefficiency. Fed. R. Evid. 403. In *Daubert*, 509 U.S. at 591, the Court invoked the concept of “fit,” which addresses the relationship of the basis for an expert’s scientific opinion to the facts of the case and the issues in dispute. In a toxic substance case in which cause in fact is disputed, an epidemiologic study of the same agent to which the plaintiff was exposed that examined the association with the same disease from which the plaintiff suffers would undoubtedly have sufficient “fit” to be a part of the basis of an expert’s opinion. The Court’s concept of “fit,” borrowed from *United States v. Downing*, 753 F.2d 1224, 1242 (3d Cir. 1985), appears equivalent to the more familiar evidentiary concept of probative value, albeit one requiring assessment of the scientific reasoning the expert used in drawing inferences from methodology or data to opinion.

285. We reiterate a point made at the outset of this section: This discussion of the use of a threshold relative risk for specific causation is not epidemiology or an inquiry an epidemiologist would undertake but an effort by courts and commentators to adapt the legal standard of proof to the available scientific evidence. See *supra* text accompanying notes 274–76.

286. Edward J. Devitt & Charles B. Blackmar, 2 Federal Jury Practice and Instruction § 71.13 (3d ed. 1977); see also *United States v. Fatico*, 458 F. Supp. 388, 403 (E.D.N.Y. 1978) (“Quantified, the preponderance standard would be 50%+ probable.”), *aff’d*, 603 F.2d 1053 (2d Cir. 1979).

287. An adherent of the frequentist school of statistics would resist this adaptation, which may explain why many epidemiologists and toxicologists also resist it. To use an epidemiologic study outcome to determine the probability of specific causation requires a shift from a frequentist approach, which involves sampling or frequency data from an empirical test, to a subjective probability about a discrete event. A frequentist might assert, after conducting a sampling test, that 60%

the relative risk or association found in a study, before an association or relative risk is used to make a statement about the probability of individual causation, the inferential judgment²⁸⁸ that the association is truly causal is required. “[A]n agent cannot be considered to cause the illness of a specific person unless it is recognized as a cause of that disease in general.”²⁸⁹ The following discussion should be read with this caveat in mind.²⁹⁰

Some courts have reasoned that when epidemiologic studies find that exposure to the agent causes an incidence in the exposed group more than twice the incidence in the unexposed group (i.e., a relative risk greater than 2.0), the probability that exposure to the agent caused a similarly situated individual’s disease is greater than 50%. These courts hold that when there is group-based evidence finding that exposure to an agent causes an incidence of disease in the exposed group that is more than twice the incidence in the unexposed group, the evidence is sufficient to satisfy the plaintiff’s burden of production and permit submission of specific causation to a jury. In such a case, the factfinder may find that it is more likely than not that the substance caused the particular plaintiff’s disease.²⁹¹ Courts, thus, have permitted expert witnesses to testify to specific causation based on the logic of the effect of a doubling of the risk.²⁹²

of the balls in an opaque container are blue. The same frequentist would resist the statement, “The probability that a single ball removed from the box and hidden behind a screen is blue is 60%.” The ball is either blue or not, and no frequentist data would permit the latter statement. “[T]here is no logically rigorous definition of what a statement of probability means with reference to an individual instance. . . .” Lee Loevinger, *On Logic and Sociology*, 32 *Jurimetrics J.* 527, 530 (1992), <https://doi.org/10.2307/1190721>; see also Steve Gold, *Causation in Toxic Torts: Burdens of Proof, Standards of Persuasion and Statistical Evidence*, 96 *Yale L. J.* 376, 382–92 (1986). Subjective probabilities about unique events are employed by those using Bayesian methodology. See Kaye, *supra* note 115, at 54–62; Kaye & Stern, *supra* note 16, “What is Bayes’ Rule?”

288. This is described in the section titled “Effect Modification” above.

289. Cole, *supra* note 94, at 10,284.

290. We emphasize this point both because it is not intuitive and because some courts have failed to appreciate the difference between an association and a causal relationship. See, e.g., *Forsyth v. Eli Lilly & Co.*, Civil No. 95–00185 ACK, 1998 U.S. Dist. LEXIS 541, at *26–*31 (D. Haw. Jan. 5, 1998). But see *Berry v. CSX Transp., Inc.*, 709 So. 2d 552, 568 (Fla. Dist. Ct. App. 1998) (“From epidemiologic studies demonstrating an association, an epidemiologist may or may not infer that a causal relationship exists.”).

291. A similar means to address probabilities in individual cases is use of the attributable risk (or attributable fraction parameter). The attributable risk is that portion of the excess risk that can be attributed to an agent, above and beyond the background risk that is due to other causes. See section titled “Attributable Risk” above. Thus, when the relative risk is greater than 2.0, the attributable fraction exceeds 50%.

292. For a comprehensive, if dated, list of cases that support proof of causation based on group studies, see *Restatement (Third) of Torts: Liability for Physical and Emotional Harm* § 28 cmt. c(4) rptrs. note (2010). The *Restatement* catalogs those courts that require a relative risk in excess of 2.0 as a threshold for sufficient proof of specific causation and those courts that recognize that even a lower relative risk than that can support specific causation, as explained below. Despite considerable disagreement on whether a relative risk of 2.0 is required or merely a taking-off point for determining

While this reasoning has a certain logic as far as it goes, there are a number of significant assumptions that require explication. These are addressed below.

A Valid Study and Risk Estimate. The propriety of this “doubling” reasoning depends on group studies identifying a genuine causal relationship and a reasonably reliable measure of the increased risk.²⁹³ This requires attention to the possibility of random error, bias, or confounding being the source of the association rather than a true causal relationship, as explained in the sections above titled “Sources of Error in Epidemiologic Studies” and “General Causation.”

Similarity of Study Participants and Plaintiff. Only if the study participants and the plaintiff are similar with respect to relevant risk factors will a risk estimate from a study or studies be valid when applied to an individual.²⁹⁴ Thus, if those exposed in a study of the risk of lung cancer from smoking have smoked half a pack of cigarettes a day for 20 years, the degree of increased incidence of lung cancer among them cannot be extrapolated to someone who smoked two packs of cigarettes for 30 years, without strong assumptions about the dose–response relationship.²⁹⁵ This principle is also applicable to risk factors for competing causes. If all of the participants in a study are participating because they were identified as

the sufficiency of the evidence on specific causation, two commentators who surveyed the cases observed that “[t]here were no clear differences in outcomes as between federal and state courts.” Russell S. Carruth & Bernard D. Goldstein, *Relative Risk Greater than Two in Proof of Causation in Toxic Tort Litigation*, 41 *Jurimetrics J.* 195, 199 (2001), <https://www.jstor.org/stable/29762698>.

293. Indeed, one commentator contends that because epidemiologic studies often are insufficiently precise to accurately measure small increases in risk, in general, studies that find a relative risk less than 2.0 should not be sufficient for causation. This concern is not with specific causation, but with general causation and the likelihood that an association less than 2.0 is noise rather than reflecting a true causal relationship. See Michael D. Green, *The Future of Proportional Liability*, in *Exploring Tort Law* (Stuart Madden ed., 2005); see also Robert F. Reynolds, *The Use of Randomized Controlled Trials for Pharmacoepidemiology*, in *Pharmacoepidemiology* 792, 794 (Brian L. Strom ed., 4th ed. 2005) (cautioning that small relative risks in observational studies can be the product of confounding); Gary Taubes, *Epidemiology Faces Its Limits*, 269 *Sci.* 164 (1995) (explaining views of several epidemiologists about a threshold relative risk of 3.0 to seriously consider a causal relationship); N.E. Breslow & N.E. Day, *Statistical Methods in Cancer Research*, in *The Analysis of Case-Control Studies* 36 (IARC Pub. No. 32, 1980) (“[r]elative risks of less than 2.0 may readily reflect some unperceived bias or confounding factor”); David A. Freedman & Philip B. Stark, *The Swine Flu Vaccine and Guillain-Barré Syndrome: A Case Study in Relative Risk and Specific Causation*, 64 *Law & Contemp. Probs.* 49, 61 (2001), <https://doi.org/10.1177/0193841X9902300603> (“If the relative risk is near 2.0, problems of bias and confounding in the underlying epidemiologic studies may be serious, perhaps intractable.”).

294. Council on Sci. Affs., Am. Med. Ass’n, *Radioepidemiological Tables*, 257 *JAMA* 806, 807 (1987) (“The basic premise of probability of causation is that individual risk can be determined from epidemiologic data for a representative population; however, the premise only holds if the individual is truly representative of the reference population.”).

295. Conversely, a risk estimate from a study that involved a greater exposure is not applicable to an individual exposed to a lower dose. See, e.g., *In re Bextra & Celebrex Mktg. Sales Prac. & Prod. Liab. Litig.*, 524 F. Supp. 2d 1166, 1175–76 (N.D. Cal. 2007) (relative risk found in studies of those who took twice the dose of others could not support an expert’s opinion of causation for the latter group); *In re Bextra & Celebrex*, No. 762000/2006, 2008 N.Y. Misc. LEXIS 720, at *29 (Sup.

having a family history of heart disease, the magnitude of risk found in a study of the effect of smoking on the risk of heart disease cannot validly be applied to an individual without such a family history.²⁹⁶

Similarly, if an individual has been differentially exposed to other risk factors from those in a study, the results of the study will not provide an accurate basis for the probability of causation for the individual.²⁹⁷ Consider a study of the effect of smoking on lung cancer among subjects who have no asbestos exposure. The relative risk of smoking in that study would not be applicable to an asbestos insulation worker.

More generally, the relative risk found in a study represents an average risk for the group. If some other factor(s) in addition to the exposure under study affect(s) the risk of the outcome of interest, and the study subjects are heterogeneous with regard to the other risk factor(s), then even among the study participants, the individual risks would be higher or lower than the study's reported relative risk.²⁹⁸ So too for a plaintiff who did not participate in the study: depending on the plaintiff's other risk factor(s), the plaintiff's individual risk from exposure could be higher or lower than the overall relative risk observed among the study participants.²⁹⁹

Ct. Jan. 7, 2008). Extrapolation is generally not possible because the shape of the dose–response curve is not known.

296. Family history in this context implies some shared genetic basis for an increased risk of heart disease. Genetics as a competing cause of disease is discussed *infra* in notes 310–17 and accompanying text.

297. See Kaye & Stern, *supra* note 16, at 536, n.180 and accompanying text (explaining that studies that use an average do not account for individual variation); David A. Freedman & Philip Stark, *The Swine Flu Vaccine and Guillain-Barré Syndrome: A Case Study in Relative Risk and Specific Causation*, 23 Evaluation Rev. 619 (1999), <https://doi.org/10.1177/0193841X9902300603> (analyzing the role that individual variation plays in determining the probability of specific causation based on the relative risk found in a study and providing a mathematical model for calculating the effect of individual variation); Mark Parascandola, *What Is Wrong with the Probability of Causation?*, 39 Jurimetrics J. 29 (1998).

298. For example, although the evidence is conflicting, some studies have found that certain variants of the NAT2 gene increase the relative risk of breast cancer for women exposed to tobacco smoke, even though studies of NAT2 variants alone or of tobacco smoke exposure alone have not found an association with breast cancer risk. See Petra Kasajova et al., *Active Cigarette Smoking and the Risk of Breast Cancer at the Level of N-acetyltransferase 2 (NAT2) Gene Polymorphisms*, 37 Tumor Biol. 7929, 7929 (2016), <https://doi.org/10.1007/s13277-015-4685-3>. If this is correct, then an epidemiologic study of tobacco smoke exposure and breast cancer that did not account for genetic variation in NAT2 would report a relative risk that did not apply either to the study participants with the high-risk variants (for whom the relative risk of exposure would be higher) or to the study participants without the high-risk variants (for whom the relative risk of exposure would be lower). If such a study hypothetically reported a relative risk of 3.5 for those exposed to tobacco smoke, quite possibly no study participant would have a 3.5-fold increased risk; those with low-risk genetic variants would have no increased risk while those with high-risk genetic variants would have more than a 3.5-fold increased risk.

299. The comment of two prominent epidemiologists on this subject is illuminating:

We cannot measure the individual risk, and assigning the average value to everyone in the category reflects nothing more than our ignorance about the determinants of lung cancer that interact with

Causation, Not Acceleration of Disease. Another assumption embedded in using the risk findings of a group study to determine the probability of causation in an individual is the assumption that the disease never would have been contracted absent exposure. Put another way, the assumption is that the agent did not merely accelerate occurrence of the disease without affecting the lifetime risk of contracting the disease. Birth defects are an example of an outcome that is not accelerated. However, for most of the chronic diseases of adulthood, it is possible, and may even be likely, that the primary avenue of effect is acceleration of disease. Yet it is not possible for epidemiologic studies to distinguish between acceleration of disease and causation of new disease. If, in fact, acceleration is involved, the relative risk from a study will understate the probability that exposure accelerated the occurrence of the disease.³⁰⁰ Of course, acceleration, if it occurs, could affect the determination of the amount of damages.³⁰¹

Agent Operates Independently. Employing a risk estimate to determine the probability of causation is not valid if the agent interacts with another cause in a way that increases disease beyond merely the sum of the increased incidence due

cigarette smoke. It is apparent from epidemiological data that some people can engage in chain smoking for many decades without developing lung cancer. Others are or will become primed by unknown circumstances and need only to add cigarette smoke to the nearly sufficient constellation of causes to initiate lung cancer. In our ignorance of these hidden causal components, the best we can do in assessing risk is to classify people according to measured causal risk indicators and then assign the average observed within a class to persons within the class.

Lash et al., *supra* note 50, at 9. See also Ofer Shpilberg et al., *The Next Stage: Molecular Epidemiology*, 50 J. Clinical Epidemiology 633, 637 (1997). [https://doi.org/10.1016/s0895-4356\(97\)00052-8](https://doi.org/10.1016/s0895-4356(97)00052-8) (“A 1.5-fold relative risk may be composed of a 5-fold risk in 10% of the population, and a 1.1-fold risk in the remaining 90%, or a 2-fold risk in 25% and a 1.1-fold for 75%, or a 1.5-fold risk for the entire population.”).

300. The short explanation is this: Suppose during the period in which instances of disease are ascertained, the agent under study accelerates occurrence of disease that would have occurred later in the period. Such cases will then not appear as excess cases of the disease in the exposed group, because they would have also occurred in the unexposed group, albeit at a later time in the period. For further discussion, see Sander Greenland & James M. Robins, *Conceptual Problems in the Definition and Interpretation of Attributable Fractions*, 128 Am. J. Epidemiology 1185 (1986), <https://doi.org/10.1093/oxfordjournals.aje.a115073>; Sander Greenland & James M. Robins, *Epidemiology, Justice, and the Probability of Causation*, 40 Jurimetrics J. 321 (2000); Sander Greenland, *Relation of Probability of Causation to Relative Risk and Doubling Dose: A Methodologic Error That Has Become a Social Problem*, 89 Am. J. Pub. Health 1166 (1999), <https://doi.org/10.2105/ajph.89.8.1166>. If acceleration occurs, then the appropriate characterization of the harm for purposes of determining damages would have to be addressed.

301. A defendant who only accelerates the occurrence of the harm of, say, chronic back pain, that would have occurred independently to the plaintiff at a later time is not liable for the same amount of damages as a defendant who causes a lifetime of chronic back pain. See David A. Fischer, *Successive Causes and the Enigma of Duplicated Harm*, 66 Tenn. L. Rev. 1127, 1127 (1999); Michael D. Green, *The Intersection of Factual Causation and Damages*, 55 DePaul L. Rev. 671 (2006).

to each agent separately. For example, the relative risk of lung cancer due to smoking is around 10, while the relative risk from asbestos exposure is approximately 5. The relative risk for someone exposed to both is not the arithmetic sum of the two relative risks (15) but closer to the product (50– to 60-fold), reflecting an interaction between the two.³⁰² Neither of the individual agents' relative risks can be employed to estimate the probability of causation in someone exposed to both asbestos and cigarette smoke.³⁰³

Other Assumptions. Additional assumptions essential to courts' risk-doubling logic have been identified. These include: (1) the agent of interest is not responsible for fatal diseases other than the disease of interest³⁰⁴ and (2) the agent does not provide a protective effect against the outcome of interest in a subpopulation of those being studied.³⁰⁵

Courts should be cognizant of the above assumptions when assessing the sufficiency of the evidence about specific causation. When one or more of these assumptions are questionable, the probability of specific causation for a given individual may be greater or less than the attributable risk derived from epidemiologic studies of an exposure and a disease.

Evidence in a given case may challenge one or more of these assumptions. Bias in a study may suggest that the study findings are inaccurate and should be estimated to be higher or lower, or even that the findings are specious—that is, they do not reflect a true causal relationship. A plaintiff may have been exposed to a dose of the agent in question that is greater or less than that to which those in the study were exposed.³⁰⁶ A plaintiff may be more or less susceptible to the

302. We use *interaction* to mean that the combined effect is other than the additive sum of each effect, which is what we would expect if the two agents operate independently. Statisticians often employ the term *interaction* in a different manner to mean that the outcome deviates from what was expected in the model specified in advance. See Jay S. Kaufman, *Interaction Reaction*, 20 *Epidemiology* 159 (2009); Sander Greenland & Kenneth J. Rothman, *Concepts of Interaction*, in Lash et al., *supra* note 49, at 329. The potential for interaction also exists with genetic risk factors. See *supra* notes 181 and 191 and accompanying text and notes 292–99.

303. See Restatement (Third) of Torts: Liability for Physical and Emotional Harm § 28 cmt. c(5) (2010); Jan Beyea & Sander Greenland, *The Importance of Specifying the Underlying Biologic Model in Estimating the Probability of Causation*, 76 *Health Physics* 269 (1999), <https://doi.org/10.1097/00004032-199903000-00008>.

304. This is because in the epidemiologic studies relied on, those deaths caused by the alternative disease process will mask the true magnitude of increased incidence of the studied disease when the study subjects die before developing the disease of interest.

305. See Sander Greenland & James M. Robins, *Epidemiology, Justice, and the Probability of Causation*, 40 *Jurimetrics J.* 321, 332–33 (2000).

306. See section titled “Dose–Response Relationship” above; see also *McManaway v. KBR, Inc.*, 852 F.3d 444, 453 (5th Cir. 2017) (“Proof that one is similarly situated to subjects in epidemiological studies must ‘include proof that the injured person was exposed to the same substance, that the exposure or dose levels were comparable to or greater than those in the studies, that the exposure occurred before the onset of injury. . . .’” (quoting *Merrell Dow Pharms., Inc. v. Havner*,

agent's toxic effects than the average of the participants in the study. Or a plaintiff may have individual factors, such as genetic risk factors, that make it less likely that exposure to the agent caused the plaintiff's disease. Similarly, an individual plaintiff may be able to rule out other known (background) causes of the disease, such as genetics or other exposures, and thereby increase the likelihood that the agent was responsible for that plaintiff's disease. Evidence of a pathological mechanism that is relevant to the cause of the plaintiff's disease may be available.³⁰⁷ A valid biomarker of the exposure, or of the exposure's effect, may be present or absent.³⁰⁸ These (and other) possibilities should be considered, as warranted by the evidence, before any attributable risk from an epidemiologic study is used to estimate the probability that the agent in question caused an individual plaintiff's disease.³⁰⁹

Recently, genetics has received increasing attention as an individual factor that could distinguish a plaintiff's own risk from the relative risk reported in epidemiologic studies. As discussed above, genetic variations may result in variable susceptibility to the toxic effect of exposure.³¹⁰ But as suggested by the preceding paragraph, genetics may also be an independent risk factor for disease. At the extreme, a plaintiff's disease may be a purely genetic disorder—the genetic

953 S.W.2d 706, 720 (Tex. 1997)); *Baker v. Chevron U.S.A. Inc.*, 533 F. App'x 509, 520 (6th Cir. 2013) ("[T]he subjects of the [expert's] cited studies generally had much higher exposures to benzene than the plaintiffs, and thus, 'there [was] simply too great an analytical gap between the data and the opinion proffered.'" (quoting *Gen. Elec. Co. v. Joiner*, 522 U.S. 136, 146 (1997)); *Ferebee v. Chevron Chem. Co.*, 736 F.2d 1529, 1536 (D.C. Cir. 1984) ("The dose–response relationship at low levels of exposure for admittedly toxic chemicals like paraquat is one of the most sharply contested questions currently being debated in the medical community.").

307. See *In re Abilify (Aripiprazole) Prods. Liab. Litig.*, 299 F. Supp. 3d 1291, 1372–73 (N.D. Fla. 2018) (finding that plaintiff's experts appropriately relied on peer-reviewed, published scientific literature analyzing the biological mechanism by which the prescription drug Abilify can cause impulse control problems to establish general causation).

308. See section titled "Genetic and Molecular Epidemiologic Studies" above and *Eaton et al.*, *supra* note 60, "Use of Biomarkers in Toxicology Exposure Assessment." For a biomarker to be valid in this context, at the time of measurement (usually during litigation, possibly during medical treatment) the biomarker must accurately reflect an earlier exposure (possibly much earlier for diseases with long latency periods).

309. See *Merrell Dow Pharms., Inc. v. Havner*, 953 S.W.2d 706, 720 (Tex. 1997); *Smith v. Wyeth-Ayerst Labs. Co.*, 278 F. Supp. 2d 684, 708–09 (W.D.N.C. 2003) (describing expert's effort to refine relative risk applicable to plaintiff based on specific risk characteristics applicable to her, albeit in an ill-explained manner); *McDarby v. Merck & Co.*, 949 A.2d 223 (N.J. Super. Ct. App. Div. 2008); *Mary Carter Andruet, Proof of Cancer Causation in Toxic Waste Litigation*, 61 S. Cal. L. Rev. 2075, 2100–04 (1988). An example of a judge sitting as factfinder and considering individual factors for a number of plaintiffs in deciding cause in fact is contained in *Allen v. United States*, 588 F. Supp. 247, 429–43 (D. Utah 1984), *rev'd on other grounds*, 816 F.2d 1417 (10th Cir. 1987). See also *Manko v. United States*, 636 F. Supp. 1419, 1437 (W.D. Mo. 1986), *aff'd*, 830 F.2d 831 (8th Cir. 1987).

310. See *supra* notes 190–200 and accompanying text.

equivalent of a signature disease.³¹¹ A number of vaccine claims have been resolved against the claimant for this reason.³¹²

Much more typically, however, especially for the complex diseases that are often the subject of litigation, variations in genes are associated with increases in risk, sometimes relatively small ones, of diseases for which other risk factors are known. With the increasing availability of genetic information, and new methods and technologies such as genome-wide association studies (GWAS) and next-generation sequencing,³¹³ more and more of these associations are being identified³¹⁴—and are being suggested in litigation as competing causes of plaintiffs’ illnesses.³¹⁵ These associations, like any epidemiologic associations, may or

311. Well-known conditions that result from genetic variations include the sickle-cell trait (see Khoury & Dorfman, *supra* note 56, at 370), phenylketonuria (see Genetic & Rare Diseases Information Center, Phenylketonuria, <https://perma.cc/NQ8D-3XRT>), and cystic fibrosis (see Steven M. Rowe et al., *A Breath of Fresh Air*, Sci. Am., Aug. 2011, at 69, 71, <https://doi.org/10.1038/scientificamerican082011-5Uq2nwWGGgRCsPlzmDVY1>). Cystic fibrosis is an example of how some genes may occur in the population in many different variant forms that cause disease of varying severity. See Christopher Wills, Exons, Introns, and Talking Genes 212–13 (1991) (“mutations are found all over the [cystic fibrosis] gene”).

312. For example, the Federal Circuit has affirmed multiple special-master rulings that a mutation in the SCN1A gene, rather than a vaccination, more likely than not caused a child’s severe myoclonic epilepsy of infancy (SMEI, also known as Dravet’s syndrome), despite claimants’ contentions that a vaccine triggered onset of the genetically caused disease. Snyder v. Sec’y of HHS, 553 Fed. App’x 994, 999–1004 (Fed. Cir. 2014) (reversing Court of Federal Claims and reinstating special master’s denial of compensation); Deribeaux v. Sec’y of HHS, 717 F.3d 1363, 1368 (Fed. Cir. 2013) (affirming denial of compensation); Stone v. Sec’y of HHS, 676 F.3d 1373, 1384 (Fed. Cir. 2012) (same). It is not uncommon for a condition’s genetic basis or a claimant’s genotype—or both—to be discovered during the pendency of litigation. See, e.g., Sanchez v. Sec’y of HHS, 809 Fed. App’x 843, 854 (Fed. Cir. 2020) (remanding case with suggestion to reopen the record “[i]n light of the speed with which medical understanding of the course of genetic disorders such as Leigh’s syndrome has been expanding”); Bowen v. E.I. Du Pont de Nemours & Co., Inc., No. CIV.A. 97C-06-194 CH, 2005 WL 1952859 (Del. Super. Ct. June 23, 2005) (during litigation, child whose parents alleged she was sickened by exposure to a fungicide was found to have a genetic mutation thought to be the cause of child’s condition), *aff’d*, 906 A.2d 787 (Del. 2006).

313. See *infra* Glossary of Terms.

314. Witte & Thomas, *supra* note 50, at 969 (“The number of GWAS undertaken has expanded rapidly since 2005.”).

315. See, e.g., Stone v. Sec’y of Health & Hum. Servs., 676 F.3d 1373, 1381–83 (Fed. Cir. 2012) (affirming, as not arbitrary and capricious, special masters’ findings that children’s seizure disorders were caused by mutations in SCN1A gene and not by vaccine in conjunction with genetic susceptibility); *In re Prempro Litig.*, 586 F.3d 547, 566 (8th Cir. 2009) (affirming admission of plaintiff’s expert testimony on specific causation and jury verdict for plaintiff where defense expert testified that genetics were the cause but “every available genetic test . . . came back negative for the most common breast cancer genes”); Rhyne v. U.S. Steel Corp., 474 F. Supp. 3d 733, 753–55 (W.D.N.C. 2020) (denying motion to exclude testimony of plaintiff’s expert witness on specific causation despite acknowledged existence of some genetically caused cases of disease and lack of genetic testing of plaintiff); Bowen v. E.I. Du Pont de Nemours & Co., No. CIV.A. 97C-06-194 CH, 2005 WL 1952859 (Del. Super. Ct. June 23, 2005), *aff’d*, 906 A.2d 787 (Del. 2006) (excluding plaintiff’s expert testimony when, while litigation was pending, a newly discovered test for a newly

may not be causal and may be affected by systematic and nonsystematic error. Moreover, modeling them as competing causes to a toxic agent requires an assumption of independence, which may or may not be justified.³¹⁶ Studies of the association of genes with disease risk greatly outnumber studies that consider the possibility that genes may increase disease risk through interactions with environmental exposures.³¹⁷

Having additional evidence (genetic or otherwise) that bears on individual causation has led a few courts to conclude that a plaintiff may satisfy his or her burden of production even if a relative risk less than 2.0 emerges from the epidemiologic evidence.³¹⁸ For example, genetics might be known to be independently responsible for 50% of the incidence of a disease independent of exposure to the agent.³¹⁹ If the risk-conferring genotype can be ruled out in an individual's case, then a relative risk greater than 1.5 might be sufficient to support an inference that the agent was more likely to be responsible for the plaintiff's disease.³²⁰

discovered genetic variant established "the cause of" plaintiff's condition); *see also* Marchant, *supra* note 154, at 24 (describing cases).

316. *See supra* text accompanying notes 294–303.

317. *See* Witte & Thomas, *supra* note 50, at 969, 971–72.

318. *In re* Hanford Nuclear Reservation Litig., 292 F.3d 1124, 1137 (9th Cir. 2002) (applying Washington law) (recognizing the role of individual factors that may modify the probability of causation based on the relative risk); *Magistrini v. One-Hour Martinizing Dry Cleaning*, 180 F. Supp. 2d 584, 606 (D.N.J. 2002) ("[A] relative risk of 2.0 is not so much a password to a finding of causation as one piece of evidence, among others for the court to consider in determining whether an expert has employed a sound methodology in reaching his or her conclusion."); *Miller v. Pfizer, Inc.*, 196 F. Supp. 2d 1062, 1079 (D. Kan. 2002) (rejecting a threshold of 2.0 for the relative risk and recognizing that even a relative risk greater than 2.0 may be insufficient); *Pafford v. Sec'y, Dep't of Health & Hum. Servs.*, 64 Fed. Cl. 19 (2005) (acknowledging that epidemiologic studies finding a relative risk of less than 2.0 can provide supporting evidence of causation), *aff'd*, 451 F.3d 1352 (Fed. Cir. 2006).

319. *See generally* Gold, *supra* note 154, at 384–89, 412–16 (discussing genetic causes of disease and the role of genetic risk factors in causation disputes); Gary E. Marchant, *Genetic Susceptibility and Biomarkers in Toxic Injury Litigation*, 41 *Jurimetrics J.* 67, 67–68, 71–72, 90 (2000), <http://dx.doi.org/10.2139/ssrn.248461> (discussing role that knowledge about genetic contribution to disease might play in refining probability of causation based on epidemiologic studies of heterogeneous populations).

320. The use of probabilities in excess of 0.50 to support a verdict results in an all-or-nothing approach to damages that some commentators have criticized. The criticism is over the fact that defendants responsible for toxic agents with a relative risk just above 2.0 may be required to pay damages not only for the disease that their agents caused, but also for all instances of the disease. Similarly, those defendants whose agents increase the risk of disease by less than a doubling may not be required to pay damages for any of the disease that their agents caused. *See, e.g.*, Am. Law Inst., 2 Reporter's Study on Enterprise Responsibility for Personal Injury: Approaches to Legal and Institutional Change 369–75 (1991). Judge Posner has been in the vanguard of those advocating that damages be awarded on a proportional basis that reflects the probability of causation or liability. *See, e.g.*, *Doll v. Brown*, 75 F.3d 1200, 1206–07 (7th Cir. 1996). But to date, courts have not adopted a rule that would apportion damages based on the probability of cause in fact in toxic substances cases. *See* Green, *supra* note 293.

This idea of eliminating a known and competing cause is central to the methodology often referred to as differential diagnosis,³²¹ but it is more accurately referred to as differential etiology. The term differential diagnosis in a clinical context refers to identifying the patient's medical condition, whereas differential etiology refers to identifying the causal factors involved in an individual's condition. This is an important distinction, because for many health conditions, the cause of the condition has no relevance to its treatment, and physicians therefore do not pursue that question.³²²

The logic of differential etiology is sound: Eliminating other known and competing causes increases the probability that a given individual's disease was caused by exposure to the agent. In a differential etiology, an expert first determines other known causes of the disease in question and then attempts to ascertain whether those competing causes can be ruled out as a cause of plaintiff's disease,³²³ as in the genetics example above. Similarly, an expert attempting to determine whether an individual's emphysema was caused by occupational chemical exposure would inquire whether the individual was a smoker. By ruling out (or ruling in) the possibility of other causes, the probability that a given agent was the cause of an individual's disease can be refined. Differential etiologies are most critical when the agent at issue is relatively weak and is not responsible for a large proportion of the disease in question.

321. Physicians regularly employ differential diagnoses in treating their patients to identify the disease from which the patient is suffering. See Jennifer R. Jamison, *Differential Diagnosis for Primary Practice* (1999).

322. See *Zandi v. Wyeth a/k/a Wyeth, Inc.*, No. 27-CV-06-6744, 2007 WL 3224242 (Minn. Dist. Ct. Oct. 15, 2007) (commenting that physicians do not attempt to determine the cause of breast cancer); see also John B. Wong et al., *Reference Guide on Medical Testimony*, "Medical Versus Legal Terminology," in this manual; Edward J. Imwinkelried, *The Admissibility and Legal Sufficiency of Testimony About Differential Diagnosis (Etiology): of Under- and Over-Estimations*, 56 *Baylor L. Rev.* 391, 402–03 (2004); *Turner v. Iowa Fire Equip. Co.*, 229 F.3d 1202, 1208 (8th Cir. 2000) (distinguishing between differential diagnosis conducted for the purpose of identifying the disease from which the patient suffers and one attempting to determine the cause of the disease); *Creanga v. Jarda*, 886 A.2d 633, 639 (N.J. 2005) ("Whereas most physicians use the term to describe the process of determining which of several *diseases* is causing a patient's *symptoms*, courts have used the term in a more general sense to describe the process by which *causes* of the patient's *condition* are identified." (quoting *Clausen v. M/V New Carissa*, 339 F.3d 1049, 1057 n.4 (9th Cir. 2003))).

323. Courts regularly affirm the legitimacy of employing differential diagnostic methodology. See, e.g., *In re Ephedra Prods. Liab. Litig.*, 393 F. Supp. 2d 181, 187 (S.D.N.Y. 2005); *Easum v. Miller*, 92 P.3d 794, 802 (Wyo. 2004) ("Most circuits have held that a reliable differential diagnosis satisfies *Daubert* and provides a valid foundation for admitting an expert opinion. The circuits reason that a differential diagnosis is a tested methodology, has been subjected to peer review/publication, does not frequently lead to incorrect results, and is generally accepted in the medical community." (quoting *Turner v. Iowa Fire Equip. Co.*, 229 F.3d 1202, 1208 (8th Cir. 2000))); *Alder v. Bayer Corp., AGFA Div.*, 61 P.3d 1068, 1084–85 (Utah 2002).

Although differential etiology is a sound methodology in principle, this approach is only valid if general causation exists³²⁴ and a substantial proportion of competing causes are known.³²⁵ But for diseases for which the causes are largely unknown, such as most birth defects, a differential etiology is of little benefit.³²⁶ And like any scientific methodology, it can be performed in an unreliable manner.³²⁷

324. “Courts almost universally agree that one must ‘rule in’ the putative cause of an injury before ruling out alternative causes.” Joseph Sanders et al., *Differential Etiology: Inferring Specific Causation in the Law from Group Data in Science*, 63 Ariz. L. Rev. 851, 874 (2021). An excellent explanation for why differential etiologies generally are inadequate without further proof of general causation was provided in *Cavallo v. Star Enterprises*, 892 F. Supp. 756 (E.D. Va. 1995), *aff’d in relevant part*, 100 F.3d 1150 (4th Cir. 1996):

The process of differential diagnosis is undoubtedly important to the question of “specific causation”. If other possible causes of an injury cannot be ruled out, or at least the probability of their contribution to causation minimized, then the “more likely than not” threshold for proving causation may not be met. But, it is also important to recognize that a fundamental assumption underlying this method is that the final, suspected “cause” remaining after this process of elimination must actually be capable of causing the injury. That is, the expert must “rule in” the suspected cause as well as “rule out” other possible causes. And, of course, expert opinion on this issue of “general causation” must be derived from a scientifically valid methodology.

Cavallo, 892 F. Supp. at 771 (footnote omitted); *see also* C.W. *ex rel.* Wood v. Textron, Inc., 807 F.3d 827, 837–38, 839 (7th Cir. 2015); *Tamraz v. Lincoln Elec. Co.*, 620 F.3d 665, 674–76 (6th Cir. 2010); *Hoefling v. U.S. Smokeless Tobacco Co., LLC*, 576 F. Supp. 3d 262, 280–85 (E.D. Pa. 2021); *see generally* Joseph Sanders & Julie Machal-Fulks, *The Admissibility of Differential Diagnosis Testimony to Prove Causation in Toxic Tort Cases: The Interplay of Adjective and Substantive Law*, 64 Law & Contemp. Probs. 107, 122–25 (2001) (discussing cases rejecting differential diagnoses in the absence of other proof of general causation and contrary cases).

325. Courts have long recognized that to prove causation, the plaintiff need not eliminate all potential competing causes, only a sufficient number to conclude that the defendant’s agent was more likely than not a cause of the plaintiff’s disease. *See Stubbs v. City of Rochester*, 134 N.E. 137, 140 (N.Y. 1919). Of course, before a competing cause should be considered relevant to a differential diagnosis, there must be adequate evidence that *it* is a cause of the disease. *See Cooper v. Smith & Nephew, Inc.*, 259 F.3d 194, 202 (4th Cir. 2001); *Ranes v. Adams Labs., Inc.*, 778 N.W.2d 677, 690 (Iowa 2010).

326. *See* *Henricksen v. Conoco Phillips Co.*, 605 F. Supp. 2d 1142, 1162 (E.D. Wash. 2009) (excluding expert’s differential diagnosis testimony because “[s]tanding alone, the presence of a known risk factor is not a sufficient basis for ruling out idiopathic origin in a particular case”); *Perry v. Novartis Pharms. Corp.*, 564 F. Supp. 2d 452, 469 (E.D. Pa. 2008) (finding experts’ testimony inadmissible because of failure to account for idiopathic causes in conducting differential diagnosis); *Soldo v. Sandoz Pharms. Corp.*, 244 F. Supp. 2d 434, 480, 519 (W.D. Pa. 2003) (criticizing expert for failing to account for idiopathic causes); *Magistrini v. One-Hour Martinizing Dry Cleaning*, 180 F. Supp. 2d 584, 609 (D.N.J. 2002) (observing that 90–95% of leukemias are of unknown causes, but proceeding incorrectly to assert that plaintiff was obliged to prove that her exposure to defendant’s benzene was the cause of her leukemia rather than simply a cause of the disease that combined with other exposures to benzene).

327. Numerous courts have concluded that, based on the manner in which a differential diagnosis was conducted, it was unreliable and the expert’s testimony based on it is inadmissible. *See, e.g., Glastetter v. Novartis Pharms. Corp.*, 252 F.3d 986, 989 (8th Cir. 2001); *see generally* Joseph

Qualifications of Experts in Epidemiology

An epidemiology expert witness must be able to identify the relevant research literature and critically review its research methods, statistical analyses, and application to the causal issue presented in the case. No single academic degree, research specialty, or career path qualifies an individual as an expert in epidemiology. Many such experts hold a doctoral degree (Ph.D., Dr.P.H., or Sc.D.) in epidemiology or a related field such as public health or biostatistics; a master's degree (M.P.H., typically) or equivalent training, education, or work experience is usually required of an expert.³²⁸ Many epidemiologists have received a medical degree plus additional training in epidemiology, often a master's degree. Epidemiologists often specialize in subfields such as the epidemiology of infectious diseases, perinatal or reproductive outcomes, chronic diseases such as cancer or cardiovascular diseases, psychosocial factors, environmental or occupational health, research methods, or molecular epidemiology. Most courts have recognized that an expert epidemiologist need not be a clinician and is qualified to provide expert opinions even if the witness has not conducted research about the agent or disease involved in the case.³²⁹

Sanders et al., *Differential Etiology: Inferring Specific Causation in the Law from Group Data in Science*, 63 Ariz. L. Rev. 851 (2021).

328. For illustrations of the types of credentials courts routinely endorse as qualifying a witness to provide expert opinions about epidemiology, including testimony about the meaning and interpretation of epidemiologic studies, see, e.g., the following: *In re Proton-Pump Inhibitor Prods. Liab. Litig.*, No. 2:17-MD-2789 (CCC) (LDW) (MDL 2789), 2022 WL 18999830, at *12–*13 (D.N.J. July 5, 2022) (Ph.D. in biostatistics although not an epidemiologist); *Holcombe v. United States*, 516 F. Supp. 3d 660, 674 (W.D. Tex. 2021) (M.P.H. with doctorate in health policy and management); *In re Johnson & Johnson Talcum Powder Prods. Mktg. Sales Prac. & Prods. Litig.*, 509 F. Supp. 3d 116, 158 (D.N.J. 2020) (Ph.D. in epidemiology with M.D. degree); *id.* at 188 (professor of environmental health sciences and epidemiology who held M.D. degree and M.H.S. in epidemiology and clinical epidemiology); *Wagoner v. Exxon Mobil Corp.*, 813 F. Supp. 2d 771, 800 (E.D. La. 2011) (hematopathologist with M.P.H. who was board-eligible in occupational and environmental medicine and had years of experience consulting on occupation-related malignancies); *In re Welding Fume Prods. Liab. Litig.*, No. 1:03-CV-17000, 2010 WL 7699456, at *29 (N.D. Ohio June 4, 2010) (Ph.D. statistician who was not an epidemiologist but taught in a clinical epidemiology department); *id.* at *40–*41 (holder of M.P.H. and medical degrees who was board-certified in internal medicine and occupational medicine and “served as a medical epidemiologist with the Centers for Disease Control”); *id.* at *45–*46 (M.D. board-certified in preventive and occupational medicine who completed fellowship in occupational neurology and held an M.P.H.); *Tyler ex rel. Tyler v. Sterling Drug, Inc.*, 19 F. Supp. 2d 1239, 1243 (N.D. Okla. 1998) (M.D. who had been director of the Office of Epidemiology within the United States Food & Drug Administration).

329. E.g., *Johnson & Johnson Talcum Powder*, *supra* note 328, 509 F. Supp. 3d at 188 (holding that expert was “qualified to testify on epidemiology” in case involving ovarian cancer, even though witness’s primary research focus was pulmonary disease); *In re Fosamax Prods. Liab. Litig.*, 645 F. Supp. 164, 206 (S.D.N.Y. 2009) (holding that epidemiologist was qualified to testify about

Individuals with medical or other health professional degrees sometimes offer expert opinions on matters related to epidemiology.³³⁰ In some cases, courts applying the standard for expert witness qualification³³¹ have found that individuals without formal degrees in epidemiology have obtained sufficient expertise through their research or other experience.³³² Yet medical expertise and epidemiologic

general causation despite lack of clinical experience treating patients); *In re Silicone Gel Breast Implants Prods. Liab. Litig.*, 318 F. Supp. 2d 879, 895 (C.D. Cal. 2004) (holding that epidemiologist was qualified to provide expert opinion about epidemiologic studies although expert had not conducted studies of the exposure at issue, did not specialize in epidemiology of the disease at issue, and was not a medical doctor); *Erickson v. Baxter Healthcare, Inc.*, 151 F. Supp. 2d 952 (N.D. Ill. 2001) (holding, in case alleging that defendant manufactured and did not warn of virus-infected blood product, that epidemiologists specializing in blood-borne diseases were qualified to testify about state of the art despite lack of experience treating patients, running a blood bank, or manufacturing blood products). *But see* *Blackwell v. Wyeth*, 971 A.2d 235 (Md. 2009) (affirming ruling that board-certified psychiatrist who held M.P.H. degree was not qualified to testify that thimerosal-containing vaccines can cause autism).

330. One reason this occurs is the overlap of scientific evidence relevant to general causation (as to which epidemiology often plays a leading role) and to specific causation (about which epidemiologic evidence may have probative value although, as we noted in the section above titled “Specific Causation,” epidemiology does not directly address specific causation). *See supra* footnotes 281–85 and accompanying text. Courts have sometimes distinguished between the qualifications needed to opine about general causation and the qualifications needed to opine on specific causation. *Compare, e.g.,* *Smith v. Pfizer Inc. (Roerig Div.)*, 2001 WL 968369 (D. Kan. Aug. 14, 2001) (holding, in case alleging that a prescription drug caused a violent outburst, that a psychiatrist was qualified to opine as to specific causation but was not qualified to opine as to general causation), *with* *Rhyne v. U.S. Steel Corp.*, 474 F. Supp. 3d 733, 750–51 (W.D.N.C. 2020) (holding that because “epidemiology focuses on the issue of general causation, not specific causation,” an epidemiologist who was “not a hematologist, oncologist, toxicologist, or medical doctor” was not qualified to give opinion about specific causation) *and* *Poosh v. Phillip Morris USA, Inc.*, 287 F.R.D. 543, 550–51 (N.D. Cal. 2012) (holding that an epidemiologist, although “plainly qualified to testify regarding statistics and . . . studies linking smoking and lung cancer,” was “not a practicing medical doctor, and at most could testify that in his opinion it is likely (based on statistics) that smoking contributed to the development of plaintiff’s lung cancer”). Holdings like those in *Rhyne* and *Poosh* do not address an epidemiologist’s expertise in the field but rather reflect a view of the role epidemiologic data play in proving specific causation.

331. Federal Rule of Evidence 702 refers to a “witness who is qualified as an expert by knowledge, skill, experience, training, or education.” *See In re Proton-Pump Inhibitor Prods. Liab. Litig.*, 2022 WL 18999830, at *4 (D.N.J. July 5, 2022) (Rule 702 “does not require that an expert possess the best formal or substantive qualifications”).

332. *See generally In re Paoli R.R. Yard PCB Litig.*, 916 F.2d 829, 855–56 (3d Cir. 1990) (“various kinds of ‘knowledge, skill, experience, training, or education’ qualify an expert as such. . . . [W]e hold that the district court abused its discretion” by rejecting experts “simply because the experts did not have the degree or training which the district court apparently thought would be most appropriate.”). For examples of courts holding that experience qualified witnesses as experts in epidemiology, *see, e.g., In re Proton-Pump Inhibitor*, No. 2:17-MD-2789 (CCC) (LDW) (MDL 2789), 2022 WL 18999830, at *20–*21 (D.N.J. July 5, 2022) (holding that medical doctor with Ph.D. in biochemistry and master’s degree in biometrics, who had worked at the United States Food & Drug Administration, was qualified to testify about analysis of preclinical- and clinical-trial data); *In re Mirena IUD Prods. Liab. Litig.*, 169 F. Supp. 3d 396, 426 (S.D.N.Y. 2016) (holding

expertise are not identical.³³³ For medical professionals, training in epidemiology through postdoctoral training or a formal degree program is a strong indicator of

that OB/GYNs who were not epidemiologists but had “experience evaluating (and in some cases conducting) epidemiological studies” were qualified “to opine on epidemiological studies”); *id.* at 477 (holding that witness who was not an epidemiologist but who “helped design and issue an epidemiological study,” who “received training from an epidemiologist,” and whose work “involved analyzing research” by drug companies was qualified to testify about epidemiologic studies); *Arias v. Dyncorp*, 928 F. Supp. 2d 10, 20 (D.D.C. 2013) (holding that physician board-certified in occupational medicine who had “vast experience in environmental medicine, conducting risk assessments, and assessing the genesis of diseases caused by toxins” qualified as an expert despite lack of “formal education in epidemiology or toxicology”); *In re Yasmin & Yaz (Drospirenone) Mktg. Sales Pracs. & Prods. Liab. Litig.*, No. 3:09-CV-10012-DRH, 2011 WL 6740363 (S.D. Ill. Dec. 22, 2011) (denying motion to exclude testimony of six medical doctors who were not epidemiologists but had experience with clinical trials or epidemiologic studies); *In re Welding Fume Prods. Liab. Litig.*, *supra* note 328, No. 1:03-CV-17000, 2010 WL 7699456, at *29 (holding that industrial hygienists who were not epidemiologists were qualified to opine about whether literature supported conclusion that manganese is a neurotoxin); *Ashburn v. Gen. Nutrition Ctr.*, 533 F. Supp. 2d 770, 773 (N.D. Ohio 2008) (holding that holder of Ph.D. in exercise physiology who had designed studies of a nutritional supplement was qualified to opine as to general causation in a case involving that supplement).

333. Thus, in *Blackwell v. Wyeth*, 971 A.2d 235 (Md. 2009), plaintiffs alleged that a thimerosal-containing vaccine had caused a child’s autism. The court held that because “the complex field of epidemiology [was] central” to the case, it was not an abuse of discretion to reject as unqualified a board-certified genetic counselor, a professor of pharmacology, and a pediatrician who lacked degrees, certification, or experience in epidemiology. *Id.* at 268. Other courts have held medical professionals unqualified as experts on epidemiology. *See, e.g., Soldo v. Sandoz Pharm. Corp.*, 244 F. Supp. 2d 434, 571 (W.D. Pa. 2003) (physician who was not an epidemiologist or statistician was not qualified to opine that drug caused stroke); *In re Welding Fume Prods. Liab. Litig.*, *supra* note 328, at *42 (N.D. Ohio 2010) (pulmonologist was not qualified to opine about causal connection between welding fume exposure and manganese-induced parkinsonism); *Loverdi v. Medifast, Inc.*, 385 F. Supp. 3d 399, 407 (E.D. Pa. 2019) (nutritionist, who held a doctorate in interdisciplinary arts and sciences with a concentration in nutritional sciences, was not qualified to opine on causation because expert was “not qualified to make medical diagnoses [and had] no expertise in epidemiology”); *Morin v. United States*, 534 F. Supp. 2d 1179, 1185 (D. Nev. 2005) (plaintiff’s treating physician, who had “no expertise in toxicology, epidemiology, risk-assessment, or environmental medicine,” was not qualified to provide expert opinion of causal link between plaintiff’s exposure to jet fuel and plaintiff’s cancer); *Sutera v. Perrier Grp. of Am. Inc.*, 986 F. Supp. 655, 667 (D. Mass. 1997) (oncologist/hematologist who reviewed but had “quite limited” familiarity with relevant epidemiologic and toxicological literature was not qualified “to render an opinion as to whether . . . exposures to low levels of benzene” caused plaintiff’s illness).

By contrast, in *In re Yasmin and Yaz*, *supra* note 324, at *5, plaintiffs asked the court to rule that six employees of the defendant were “not qualified to comment on the design, methodology, or reliability of the epidemiological studies at issue.” All six witnesses had medical degrees. All admitted (in varying language) that they were not epidemiologists and did not have degrees or certifications in epidemiology, public health, or biostatistics. But all had experience with some combination of reviewing, evaluating, presenting data from, conducting, or designing clinical trials or other epidemiologic studies. The court denied plaintiff’s motion in limine with leave to renew after voir dire of each witness. *Id.* at *6–*10. Other courts have held medical professionals who were not epidemiologists qualified to give expert opinions on epidemiologic matters. *See, e.g., Miller v. Bayer Healthcare Pharm. Inc.*, 2016 WL 9047163, at *2 (W.D. Mo. Dec. 20, 2016) (“although the

expertise in the field. Organizations such as the American Board of Medical Specialties offer residencies in preventive medicine to physicians where they may receive training in epidemiology.³³⁴ Belonging to epidemiologic professional societies, such as the International Society of Environmental Epidemiology (ISEE) or the Society for Epidemiologic Research (SER), may demonstrate a commitment to the field and provide additional opportunities for continuing education. Courts do not necessarily require medical professionals with epidemiologic experience to possess such credentials to qualify as experts, however.³³⁵

Often an epidemiology expert has made significant contributions to research, including, but not limited to, experience conducting research studies, publishing papers in peer-reviewed journals, leading research teams, and mentoring others in the field. Such an expert may be recognized by peers with professional distinctions in the form of awards or invitations to speak at conferences. These accomplishments and recognitions indicate that an epidemiologist is acknowledged as an expert by other epidemiologists, well beyond the threshold to testify as an expert witness.

experts are not epidemiologists . . . the court finds that [four medical doctors] have sufficient knowledge and experience to qualify them” as expert witnesses); *Wagoner v. Exxon Mobil Corp.*, 813 F. Supp. 2d 771, 800 (E.D. La. 2011) (holding that treating physician who “had maintained a substantial interest in epidemiology . . . directly tied to his work as a treating physician . . . that has led him to become familiar with epidemiological studies” was qualified to opine as to general causation); *In re Vioxx Prods. Liab. Litig.*, 401 F. Supp. 2d 565, 590–92 (E.D. La. 2005) (holding, over objection that epidemiology expertise was required, that M.D. pathologists could rely on scientific literature to testify that drug caused decedent’s death); *Burton v. R.J. Reynolds Tobacco Co.*, 183 F. Supp. 2d 1308, 1311–12 (D. Kan. 2002) (holding that physician board-certified in internal medicine and pulmonary medicine, who had written numerous articles about health effects of smoking, was qualified to opine that smoking causes peripheral vascular disease (PVD), even though witness was not an epidemiologist and had never conducted a study of PVD); *id.* at 1314 (holding that plaintiff’s treating physician, a specialist in rehabilitation therapy, was qualified to testify that smoking caused plaintiff’s PVD); *Burton v. R.J. Reynolds Tobacco Co.*, 181 F. Supp. 2d 1256, 1267 (D. Kan. 2002) (concluding that because general causation was “within the reasonable confines of” vascular surgeon’s subject area, surgeon was qualified to opine that cigarette smoking caused plaintiff’s PVD “so long as it is established at trial that [the expert’s] opinions and assessments flow from his education and experience as a vascular surgeon”); *In re Joint E. & S. Dist. Asbestos Litig.*, 52 F.3d 1124, 1136 n.21 (2d Cir. 1995) (noting that trial court should have allowed plaintiff’s treating physician, a specialist in internal medicine, to give differential diagnosis testimony, although witness was not expert in asbestos-related illness or familiar with epidemiologic literature).

334. See *Knight v. Kirby Inland Marine, Inc.*, 363 F. Supp. 2d 859, 863 (N.D. Miss. 2005) (holding that an M.D. with certification from American Board of Preventive Medicine/Occupational Medicine, who had written articles on environmental and occupational health, was “a highly qualified epidemiologist and physician”).

335. See *Wagoner v. Exxon Mobil Corp.*, 813 F. Supp. 2d 771, 800 (E.D. La. 2011) (holding that a treating physician who “had maintained a substantial interest in epidemiology” qualified as an expert witness in the field); *Wolfe v. McNeil-PPC, Inc.*, 881 F. Supp. 2d 650, 659 (E.D. Pa. 2012) (holding that a witness’s “extensive medical training and experience” were sufficient to qualify the witness to interpret epidemiologic studies although the witness was not an epidemiologist).

Acknowledgments

The authors are grateful for the able research assistance provided by Murphy Horne (Wake Forest Law School Class of 2012) and Cory Randolph (Wake Forest Law School Class of 2010) in the preparation of the third edition of this reference guide. The authors are grateful for the able research assistance provided by Lyndsey Arneson Field (Wake Forest Law School Class of 2024) and Tia Mitchell, Willa Sweeney, Elizabeth van Winkle, and Laura Ensminger (Rutgers Law School Class of 2024) in the preparation of this fourth edition of the guide. The authors also acknowledge the kind assistance of Eli M. Snir, Senior Lecturer in Data Analytics, Olin Business School, Washington University in St. Louis, in constructing the diagram of a normal distribution and the dose–response curves contained in this reference guide.

Glossary of Terms

The following terms and definitions were adapted from a variety of sources, including *A Dictionary of Epidemiology* (Miquel M. Porta et al. eds., 6th ed. 2014); Joseph L. Gastwirth, *Statistical Reasoning in Law and Public Policy*, Vol. 1 (1988); James K. Brewer, *Everything You Always Wanted to Know About Statistics, But Didn't Know How to Ask* (1978); R.A. Fisher, *Statistical Methods for Research Workers* (1973); National Human Genome Research Institute, Talking Glossary of Genomic and Genetic Terms, <https://perma.cc/43TP-QECX>; National Cancer Institute, NCI Dictionary of Cancer Terms, <https://www.cancer.gov/publications/dictionaries/cancer-terms> (last visited Mar. 16, 2025); National Cancer Institute, NCI Dictionary of Genetics Terms, <https://www.cancer.gov/publications/dictionaries/genetics-dictionary> (last visited Mar. 16, 2025).

adjustment. Methods of accounting for factors that may distort the estimated association between an exposure being studied and a disease outcome. See also direct standardization, indirect standardization.

agent. Also called risk factor. A factor, such as a drug, microorganism, chemical substance, or form of radiation, whose presence or absence can result in the occurrence of a disease. A disease may be caused by a single agent or a number of independent alternative agents, or the combined presence of a complex of two or more factors may be necessary for the development of the disease.

allele. One of two or more forms (variants) of a gene. Individuals ordinarily inherit two copies of each gene, one from each parent. The alleles inherited from each parent may be the same or different.

alpha. The level of statistical significance chosen by a researcher to determine if any association found in a study is sufficiently unlikely to have occurred by chance (as a result of random sampling error) if the null hypothesis (no association) is true. Researchers commonly adopt an alpha of 0.05, but the choice is arbitrary, and other values can be justified.

alpha error. Also called Type I error and false-positive error, alpha error occurs when a researcher rejects a null hypothesis when it is actually true (i.e., when there is no association). This can occur when an apparent difference is observed between the control group and the exposed group, but the difference is not real (i.e., it occurred by chance). A common error made by lawyers, judges, and academics is to equate the level of alpha with the legal burden of proof.

association. The degree of statistical relationship between two or more events or variables. Events are said to be associated when they occur more or less frequently together than one would expect by chance. Association does not necessarily imply a causal relationship. Events are said not to have an

association when the agent (or independent variable) does not appear to be related to the incidence of a disease (the dependent variable). This corresponds to a relative risk of 1.0 or to a beta coefficient of 0. A negative association means that the events occur less frequently together than one would expect by chance, thereby implying a preventive or protective role for the agent (e.g., a vaccine).

attributable risk among the exposed. Also called attributable fraction and attributable proportion among the exposed. The proportion of disease in exposed individuals that can be attributed to exposure to an agent, as distinguished from the proportion of disease attributed to all other causes. Compare population attributable risk.

background rate of disease. Also called background risk of disease. Rate of disease in a population that has no known exposures to an alleged risk factor for the disease. For example, the background risk for all birth defects is 3–5% of live births.

Bayesian analysis. A method of statistical inference that involves using Bayes' theorem. It allows for combining prior information about a hypothesis with new evidence to get a revised estimate of the likelihood of the hypothesis. As opposed to standard (also called frequentist) analysis, which generates confidence intervals, Bayesian methods can be used to estimate so-called credible intervals which represent intervals that have a particular probability of containing the true value of an association given the observed data. Hence, if a study was repeated multiple times under identical conditions, a 95% credible interval would be expected to contain the true value 95% of the time.

beta coefficient. A measure of the strength of the association between an exposure and an outcome. The beta coefficient is calculated as the mean change in the outcome per unit change in the exposure. For example, if a study found that the beta coefficient for the association between prenatal exposure to lead (e.g., determined based on maternal serum concentrations expressed in micrograms per deciliter [$\mu\text{g}/\text{dL}$]) and child IQ was -0.3 , this would mean that, on average, child IQ would be 0.3 IQ point lower for every $\mu\text{g}/\text{dL}$ increase in maternal serum lead concentration.

beta error. Also called Type II error and false-negative error. Occurs when a researcher fails to reject a null hypothesis when it is incorrect (i.e., when there is an association). This can occur when no statistically significant difference is detected between the control group and the exposed group, but a difference does exist.

bias. Any effect at any stage of investigation or inference tending to produce results that depart systematically from the true values. In epidemiology, the term bias does not necessarily carry an imputation of prejudice or other

subjective factor, such as the experimenter's desire for a particular outcome. This differs from conventional usage, in which bias refers to a partisan point of view.

biological marker. See biomarker.

biological plausibility. Consideration of existing knowledge about human biology and disease pathology to provide a judgment about the plausibility that an agent causes a disease.

biomarker. Also called biological marker. A physiological change in tissue or body fluids that occurs as a result of an exposure to an agent and that can be detected in the laboratory. Biological markers are only available for a small number of chemicals.

case-comparison study. See case-control study.

case-control study. Also case-comparison study, case history study, case referent study, retrospective study. A study that starts with the identification of persons with a disease (or other outcome variable) and a suitable control (comparison, reference) group of persons without the disease. Such a study is sometimes referred to as retrospective because it starts after the onset of disease and looks back to the hypothesized causal factors. However, cohort studies can also use a retrospective study design. Thus, case-control study is not synonymous with retrospective study.

case group. A group of individuals who have been exposed to the disease, intervention, procedure, or other variable being studied for its influence on subjects.

causation. As used here, an event, condition, characteristic, or agent being a necessary element of a set of other events that can produce an outcome, such as a disease. Other sets of events may also cause the disease. For example, smoking is a necessary element of a set of events that result in lung cancer, yet there are other sets of events (without smoking) that cause lung cancer. A cause may be thought of as a necessary link in at least one causal chain that results in an outcome of interest. Epidemiologists generally speak of causation in a group context; hence, they will inquire whether an increased incidence of a disease in a cohort was "caused" by exposure to an agent.

clinical trial. Also called randomized trial. An experimental study performed to assess the efficacy and safety of a drug or other treatment. Unlike observational studies, clinical trials can be conducted as experiments and use randomization, because the agent being studied is thought to be beneficial.

cohort. Any designated group of persons followed or traced over a period of time to examine health or mortality experience.

cohort study. The method of epidemiologic study in which groups of individuals can be identified who are, have been, or in the future may be

differentially exposed to an agent or agents hypothesized to influence the occurrence of a disease or other outcome. The groups are observed to determine if the exposed group is more likely to develop disease. Alternative terms for a cohort study (concurrent study, follow-up study, incidence study, longitudinal study, prospective study) describe a frequent feature of this study design, which is observation of the population for a sufficient number of person-years to generate reliable incidence or mortality rates in the population subsets. This generally implies study of a large population, study for a prolonged period (years), or both. Some cohort studies can be retrospective, determining exposure by obtaining historical information about subjects' exposure, as was the case with the path-breaking work on asbestos exposure by Irving Selikoff and collaborators in the 1960s.

confidence interval. A range of values that reflects random error. Thus, if a confidence level of 0.95 is selected for a study, 95% of similar studies would result in the true relative risk falling within the confidence interval. The width of the confidence interval provides an indication of the precision of the point estimate or relative risk found in the study. Where the confidence interval contains a relative risk of 1.0, the results of the study are not statistically significant.

confounding factor. Also called confounder. A factor that is both a risk factor for the disease and a factor associated with the exposure of interest. Confounding refers to a situation in which an association between an exposure and outcome is distorted by a factor that affects the outcome but is unaffected by the exposure.

control group. A comparison group comprising individuals who have not been exposed to the disease, intervention, procedure, or other variable whose influence is being studied.

Cox proportional-hazards model. A statistical method used to investigate the time to an event, such as death or disease, after exposure to one or more independent variables. This method might be used, for example, to determine the difference in survival time for two groups, one of which was exposed to nuclear radiation fallout.

cross-sectional study. A study that examines the relationship between disease and variables of interest as they exist in a population at a given time. A cross-sectional study measures the presence or absence of disease and other variables in each member of the study population. The data are analyzed to determine if there is a relationship between the existence of the variables and disease. Because cross-sectional studies examine only a particular moment in time, they reflect the prevalence (existence) rather than the incidence (rate) of disease and can offer only a limited view of the causal association between the variables and disease.

data dredging. Jargon that refers to results identified by researchers who, after completing a study, pore through their data seeking to find any associations that may exist. In general, good research practice is to identify the hypotheses to be investigated in advance of the study; hence, data dredging is generally frowned on. In some cases, however, researchers use this approach to conduct exploratory studies designed to generate hypotheses for further study.

demographic study. See ecological study.

deoxyribonucleic acid (DNA). A long-chain molecule that carries genetic information via the sequential arrangement of four nucleotide bases in the DNA molecule.

dependent variable. The outcome that is being assessed in a study based on the effect of another characteristic—the independent variable. Epidemiologic studies attempt to determine whether there is an association between the independent variable (exposure) and the dependent variable (incidence or prevalence of disease).

differential misclassification. A form of bias that is due to the misclassification of individuals or a variable of interest when the misclassification varies among study groups, generally with respect to the exposure or outcome under study. This type of bias occurs when, for example, the probability of incorrectly determining that individuals in a study are exposed is larger among participants with the outcome (e.g., disease) relative to those who are healthy. See nondifferential misclassification.

direct standardization. A technique used to eliminate any difference between two study populations based on age, sex, or some other parameter that might result in confounding. Direct adjustment entails comparison of the study group with a large reference population to determine the expected rates based on the characteristic, such as age, for which adjustment is being performed. See adjustment.

directed acyclic graphs (DAGs). These graphs represent researchers' assumptions about the causal structure underlying a research question as informed by scientific knowledge. As such, DAGs may vary between investigators and over time as knowledge evolves. In DAGs, variables are represented by so-called nodes, and causal connections are shown by arrows pointing from causes to effects. DAGs are acyclic because no directed path (i.e., following a series of arrows) can form a closed loop (because a variable cannot cause itself). These graphs allow researchers to identify which variable to adjust for (and which not to adjust for) in order to remove bias.

dose. Generally refers to the intensity or magnitude of exposure to an agent, taking into account the amount or concentration of the agent and the duration or frequency of exposure. In human epidemiologic studies or experimental animal models, dose typically means the amount of chemical or

physical agent that is absorbed into the body and reaches the affected tissue, not just the amount or concentration to which a person or animal is externally exposed. Some toxicologists may also refer to the external exposure as an external dose.

dose–response relationship. A relationship in which a change in amount, intensity, or duration of exposure to an agent is associated with a change—either an increase or a decrease—in risk of disease.

double-blinding. A method used in experimental studies, in which neither the individuals being studied nor the researchers know during the study whether any individual has been assigned to the exposed or control group. Double-blinding is designed to prevent knowledge of the group to which the individual was assigned from biasing the outcome of the study.

ecological fallacy. Also called aggregation bias and ecological bias. An error that occurs from inferring that a relationship that exists for groups is also true for individuals. For example, if a country with a higher proportion of fishermen also has a higher rate of suicides, then inferring that fishermen must be more likely to commit suicide is an ecological fallacy.

ecological study. Also called demographic study. A study of the occurrence of disease based on data from populations, rather than from individuals. An ecological study searches for associations between the incidence or prevalence of disease and suspected disease-causing agents in the studied populations. Researchers often conduct ecological studies by examining easily available health statistics, making these studies relatively inexpensive in comparison with studies that measure disease and exposure to agents on an individual basis.

epidemiology. The study of the distribution and determinants of disease or other health-related states and events in populations. Epidemiology may be applied to control health problems and to improve public health.

epigenetics or epigenomics. The study of biochemical constituents that regulate the degree to which genes are expressed. Epigenetic factors alter the function of the genome without altering the sequence of bases in the DNA. Certain epigenetic changes may be inherited.

epigenome. The set of chemical compounds or “marks” that provide instructions regulating gene expression in a cell.

error. See random error.

etiologic factor. An agent that plays a role in causing a disease.

etiology. The cause of disease or other outcome of interest. Also the study of the cause of disease or other outcome of interest.

experimental study. A study in which the researcher directly controls the conditions. Experimental epidemiology studies (also clinical studies) entail

random assignment of participants to the exposed and control groups (or some other method of assignment designed to minimize differences between the groups).

exposed, exposure. In epidemiology, the exposed group (or the exposed) is used to describe a group whose members have been exposed to an agent that may be a cause of a disease or health effect of interest, or possess a characteristic that is a determinant of a health outcome.

false-negative error. See beta error.

false-positive error. See alpha error.

follow-up study. See cohort study.

gene. A segment of DNA that codes for the amino acid sequence of a protein or part of a protein.

gene expression. The transcription of a DNA base sequence into RNA and translation of the RNA base sequence into an amino acid chain that forms all or part of a protein. Which genes are expressed, and how much, varies from cell to cell depending on regulatory factors that turn genes on and off.

general causation. Issue of whether an agent increases the incidence of disease in a group and not whether the agent caused any given individual's disease. Because of individual variation, a toxic agent generally will not cause disease in every exposed individual.

generalizable. When the results of a study are applicable to populations other than the study population, such as the general population.

genetic epidemiology. Epidemiology that applies genetic data to the study of health at a population level. Genetic epidemiology includes study of the genetic basis of variable phenotypes in populations, the genetic components of diseases or other health outcomes, and the ways in which genetic variations and environmental exposures or other risk factors modify each other's effects.

genetic variant. See variant.

genome. The entire set of DNA instructions found in a cell.

genome-wide association study (GWAS). A research method that involves studying the genomes of many people to attempt to identify genetic variants that are associated with increased risk of a disease or increased occurrence of some other trait (phenotype). Genes identified in a GWAS may or may not be causally associated with the trait of interest.

genomics. The study of the entire genome.

genotype. An individual's genetic code with respect to a trait of interest. For most genes, an individual's genotype ordinarily is determined by the DNA sequence found in two copies of the gene, one inherited from each biological parent.

germline or **germ line**. Of or related to the reproductive (egg or sperm) cells. A germline mutation will be inherited and present in all cells of the offspring produced by the reproductive cell carrying that mutation.

hazard ratio. The ratio of hazard rates. Hazard rates are defined as the risk of an outcome at a specific point in time. The hazard ratio is used as an estimate of the relative risk. For example, the hazard ratio might be used to describe the ratio in death rates of an exposed and an unexposed group one year after a study began.

in vitro. Within an artificial environment, such as a test tube (e.g., the cultivation of tissue in vitro). Often refers to an experiment. Compare in vivo.

in vivo. Within a living organism (e.g., the examination of an experimental animal's organ in vivo). Often refers to an experiment. Compare in vitro.

incidence rate. The number of people in a specified population falling ill from a particular disease during a given time period. More generally, the number of new events (e.g., new cases of a disease in a defined population) within a specified period of time. In an epidemiologic study, calculations of incidence rate take into account the length of time that data are collected about the study subjects.

incidence study. See cohort study.

independent variable. A characteristic that is measured in a study and that is suspected to have an effect on the outcome of interest (the dependent variable). Thus, exposure to an agent is measured in a cohort study to determine whether that independent variable has an effect on the incidence of disease, which is the dependent variable.

indirect adjustment. A technique employed to minimize error that might result when comparing two populations because of differences in age, sex, or another parameter that may independently affect the rate of disease in the populations. The incidence of disease in a large reference population, such as all residents of a country, is calculated for each subpopulation (based on the relevant parameter, such as age). Those incidence rates are then applied to the study population with its distribution of persons to determine the overall incidence rate for the study population, which provides a standardized mortality or morbidity ratio (often referred to as SMR).

inference. The intellectual process of making generalizations from observations. In statistics, the development of generalizations from sample data, usually with calculated degrees of uncertainty.

information bias. Also called observational bias. Systematic error in measuring data that results in an incorrect estimate of the relation between an exposure and an outcome.

instrumental variable. A variable or characteristic that is robustly associated with an exposure of interest and can therefore be used as a proxy to explore the unconfounded causal effect of the exposure on an outcome of interest, provided certain assumptions are satisfied. Genetic variations are used as instrumental variables in Mendelian randomization studies.

interaction. A situation in which the magnitude and/or direction (positive or negative) of the effect of one exposure differs depending on the presence or level of the other. In interaction, the effect of two exposures together is different (greater or less) than the sum of their individual effects.

inverse probability weighting (IPW). A statistical method that creates a pseudo-population by assigning weights to individuals or observations to correct for bias. It is used to correct for selection bias by generating a pseudo-population in which variables assumed to determine selection in a study sample are randomized with respect to selection. It involves weighting each individual or observation by the inverse of its selection probability to correct for being over- or underrepresented in the dataset. IPW can also be used to remove confounding by randomizing confounders with respect to exposure.

linear regression. A statistical method employed to estimate the association between one or multiple exposures and a continuous outcome (e.g., IQ). When multiple variables are included (such as more than one exposure or exposure(s) and confounder(s)), this technique is known as multiple linear regression.

logistic regression. A statistical method employed to estimate the association between one or multiple exposures and a dichotomous outcome (e.g., the existence of cancer). When multiple variables are included (such as more than one exposure or exposure(s) and confounder(s)), this technique is known as multiple logistic regression.

Mendelian randomization. An analytical tool that uses genetic variants as instrumental variables—proxies of an exposure—to assess whether an observed association between the exposure and an outcome of interest may be causal.

meta-analysis. A technique used to combine the results of several studies to enhance the precision of the estimate of the effect size and reduce the plausibility that the association found is due to random sampling error. Meta-analysis is best suited to pooling results from randomly controlled experimental studies, but if carefully performed, it also may be useful for observational studies.

misclassification bias. Bias in the estimate of an association between an exposure and an outcome due to the erroneous classification of an individual as exposed to the agent when the individual was not (or vice-versa), or

incorrectly classifying a study individual with regard to disease. Misclassification bias may be equal in all study groups (see nondifferential misclassification, sometimes referred to as random misclassification) or may vary among groups (see differential misclassification).

molecular epidemiology. Epidemiology that applies the understanding of disease at a molecular level to the study of health at a population level. Molecular epidemiology is characterized by the use of biomarker information.

morbidity rate. State of illness or disease. Morbidity rate may refer to either the incidence rate or prevalence rate of disease.

mortality rate. Proportion of a population that dies of a disease or of all causes per unit of time. The numerator is the number of individuals dying; the denominator is the total population in which the deaths occurred. The unit of time is usually a calendar year.

multivariable analysis. A set of techniques used when the variation in several variables has to be studied simultaneously. In statistics, any analytical method that allows the simultaneous study of two or more exposures or variables.

mutation. A change in the DNA sequence. Germline mutations are inherited by offspring. Somatic mutations are not inherited by offspring.

nondifferential misclassification. Error due to misclassification of individuals or of a variable of interest into the wrong category when the misclassification varies among study groups. The error may result from limitations in data collection, may result in bias, and will often produce an underestimate of the true association. See differential misclassification.

null hypothesis. A hypothesis that states that there is no true association between a variable and an outcome. At the outset of any observational or experimental study, the researcher must state a proposition that will be tested in the study. In epidemiology, this proposition typically addresses the existence of an association between an agent and a disease. Most often, the null hypothesis is a statement that exposure to Agent A does not increase the occurrence of Disease D. The results of the study may justify a conclusion that the null hypothesis (no association) has been disproved (e.g., a study that finds a strong association between smoking and lung cancer). A study may fail to disprove the null hypothesis, but that alone does not justify a conclusion that the null hypothesis has been proved.

observational study. An epidemiologic study in situations in which nature is allowed to take its course, without intervention from the investigator.

odds. For a discrete event, the ratio of the probability that the event exists to the probability that the event does not exist. Thus, if a baseball team wins half of its games, the odds of it winning are 1 to 1 or 1.0.

odds ratio (OR). Also called cross-product ratio and relative odds. The ratio of the odds that a case (one with the disease) was exposed to an agent to the odds that a control (one without the disease) was exposed to the same agent. For rare outcomes, the odds ratio from a case-control study is similar to that of the risk ratio.

p (probability) or p -value. The p -value is the probability of getting a value of the test outcome equal to or more extreme than the result observed, given that the null hypothesis is true. The letter p , followed by the abbreviation “n.s.” (not significant) means that $p > 0.05$ and that the association was not statistically significant at the 0.05 level of significance. The statement “ $p < 0.05$ ” means that p is less than 5%, and, by convention, the result is deemed statistically significant. Other significance levels can be adopted, such as 0.01 or 0.1. The lower the p -value, the less likely that random error would have produced the observed relative risk if the true relative risk is 1. See significance level; statistical significance.

parametric g-formula. A method of standardization that can be used to address confounding in causal inference with observational data. This approach estimates the impact on health outcomes of hypothetical exposure interventions.

pathognomonic. When an agent must be present for a disease to occur. Thus, asbestos is a pathognomonic agent for asbestosis. See signature disease.

phenotype. An observable trait of an individual. A phenotype of interest may result from an individual’s genotype or from a combination of genotype and environmental factors.

placebo-controlled. In an experimental study, providing an inert substance to the control group, so as to keep the control and exposed groups ignorant of their status.

polymorphic gene. A gene that exists in more than one form in the population. See polymorphism.

polymorphism. Any of multiple forms of a gene that exist in a population. Polymorphisms originally arise via mutations. The simplest polymorphism, a “single nucleotide polymorphism” or SNP, is the substitution of one particular DNA nucleotide base with another. Many millions of these have been identified. Other types of polymorphisms exist. Polymorphisms may be harmful, beneficial, or may have no discernible effect. The term polymorphism may also refer to the condition of a gene existing in multiple forms. See also variant.

population attributable risk (PAR). Also called population attributable fraction. The fraction of risk that is attributable to exposure to a substance in a population (e.g., X percent of lung cancer is attributable to cigarettes). See attributable risk among the exposed.

power. The probability that a difference of a specified amount will be detected by the statistical hypothesis test, given that a difference exists. In less formal terms, power is like the strength of a magnifying lens in its capability to identify an association that truly exists. Power is equivalent to one minus Type II error. This is sometimes stated as $Power = 1 - \beta$.

prevalence. The percentage of persons with a disease in a population at a specific point in time.

prospective study. A study in which two groups of individuals are identified: (1) individuals who have been exposed to a risk factor and (2) individuals who have not been exposed. Both groups are followed for a specified length of time, and the proportion that develops disease in the first group is compared with the proportion that develops disease in the second group. See cohort study.

proteome. The complete set of proteins made by an organism.

proteomics. The study of the proteome.

random. The term implies that an event is governed by chance. See randomization.

random error. Also called sampling error and random sampling error. Random error is the error that is due to chance when the result obtained for a sample differs from the result that would be obtained if the entire population (or universe) were studied.

randomization. Assignment of individuals to groups (e.g., for experimental and control regimens) by chance. Within the limits of chance variation, randomization should make the control group and experimental group similar at the start of an investigation and ensure that personal judgment and prejudices of the investigator do not influence assignment. Randomization should not be confused with haphazard assignment. Random assignment follows a predetermined plan that usually is devised with the aid of a table of random numbers. Randomization cannot ethically be used where the exposure is known to cause harm (e.g., cigarette smoking).

randomized trial. See clinical trial.

recall bias. Systematic error resulting from differences between two groups in a study in accuracy of memory. For example, subjects who have a disease may recall exposure to an agent more frequently than subjects who do not have the disease.

relative risk (RR). The ratio of the risk of disease or death among people exposed to an agent to the risk among the unexposed. For instance, if 10% of all people exposed to a chemical develop a disease, compared with 5% of people who are not exposed, the disease occurs twice as frequently among the exposed people. The relative risk is $10\% \div 5\% = 2$. A relative risk of 1 indicates no association between exposure and disease.

research design. The procedures and methods, predetermined by an investigator, to be adhered to in conducting a research project.

retrospective study. A research design in which subjects' exposure to a suspected toxicant or toxicants is obtained from historical information, such as medical records. Most case-control studies are retrospective studies. Some cohort studies may use a retrospective design.

ribonucleic acid (RNA). RNA exists in different forms that have different functions in cells. In the process of gene expression, messenger RNA transcribes a gene's sequence of DNA bases into an RNA sequence. The transcribed sequence is then translated into an amino acid chain with the assistance of transfer RNA.

risk. A probability that an event will occur (e.g., that an individual will become ill or die within a stated period of time or by a certain age).

risk difference (RD). The difference between the risk of disease in the exposed population and the risk of disease in the unexposed population. The value of a risk difference will lie in the range between -1 and 1. Thus, if risk in the exposed population is .25 and risk in the unexposed population is .10, the risk difference is .15.

sample. A subset of a population, which may be selected randomly or nonrandomly. Sample may also refer to an environmental or biological specimen collected to test for the presence of an agent, a disease, or a biomarker.

sample size. The number of subjects who participate in a study. Sample size may also refer to the number of environmental or biological specimens included in a study.

sampling error. See random error.

secular-trend study. Also called timeline study. A study that examines changes over a period of time, generally years or decades. Examples include the decline of tuberculosis mortality and the rise, followed by a decline, in coronary heart disease mortality in the United States in the past 50 years.

selection bias. Systematic error that results from differences between the individuals included in a study (the study sample) and the "target" population to which researchers want to infer results. Selection bias can be due to the initial selection of participants in a study, when an analysis is restricted to certain participants, when participants are lost to follow up in a cohort study, or when controls are not appropriately selected in a case-control study.

sensitivity. Measure of the accuracy of a diagnostic or screening test or device in identifying disease (or some other outcome) when it truly exists. For example, assume that we know that 20 women in a group of 1,000 women have cervical cancer. If the entire group of 1,000 women is tested for cervical cancer and the screening test only identifies 15 (of the known 20) cases

of cervical cancer, the screening test has a sensitivity of 15/20, or 75%. See also specificity.

signature disease. A disease that is associated uniquely with exposure to an agent (e.g., asbestosis and exposure to asbestos). See also pathognomonic.

significance level. A somewhat arbitrary level selected to minimize the risk that an erroneous positive study outcome that is due to random error will be accepted as a true association. The lower the significance level selected, the less likely that false-positive error will occur. See *p*-value; statistical significance.

somatic. Of or relating to a body cell other than a reproductive (sperm or egg) cell. A somatic mutation will be inherited by all cells descended from the cell in which the mutation arose but will not be found in the remainder of an organism's cells. Somatic mutations are not inherited by offspring. Cancer cells typically have many somatic mutations. Compare germline.

specific causation. Whether exposure to an agent was responsible for a given individual's disease.

specificity. Measure of the accuracy of a diagnostic or screening test in identifying those who are disease-free. Once again, assume that 980 out of a group of 1,000 women do not have cervical cancer. If the entire group of 1,000 women is screened for cervical cancer and the screening test only identifies 900 women without cervical cancer, the screening test has a specificity of 900/980, or 92%.

standardized morbidity ratio (SMR). The ratio of the frequency of disease observed in the study population to the frequency of disease that would be expected if the study population had the same strata-specific (e.g., age-specific, sex-specific) frequency of disease as some selected reference population.

standardized mortality ratio (SMR). The ratio of the frequency of death observed in the study population to the frequency of death that would be expected if the study population had the same strata-specific (e.g., age-specific, sex-specific) frequency of death as some selected reference population.

statistical significance. A term used to describe a study result or difference that exceeds the Type I error rate (or *p*-value) that was selected by the researcher at the outset of the study. In formal significance testing, a statistically significant result is unlikely to be the result of random sampling error and justifies rejection of the null hypothesis. Some epidemiologists believe that formal significance testing is inferior to using a confidence interval to express the results of a study. Statistical significance, which addresses the role of random sampling error in producing the results found in the study, should not be confused with the importance (for public health or public policy) of a research finding. See *p*-value; significance level.

stratification. Separating a group into subgroups based on specified criteria, such as age, gender, or socioeconomic status. Stratification is used both to control for the possibility of confounding (by separating the studied populations based on the suspected confounding factor) and when there are other known factors that affect the disease under study. For example, the incidence of death increases with age, and a study of mortality might use stratification of the cohort and control groups based on age.

study design. See research design.

systematic error. See bias.

teratogen. An agent that produces abnormalities in the embryo or fetus by disturbing maternal or paternal health or by acting directly on the fetus in utero.

teratogenicity. The capacity for an agent to produce abnormalities in the embryo or fetus.

threshold phenomenon. A certain level of exposure to an agent below which disease does not occur and above which disease does occur.

timeline study. See secular-trend study.

toxic substance. A substance that is poisonous.

toxicology. The science of the nature and effects of poisons. Toxicologists study adverse health effects of agents on biological organisms, such as live animals and cells. Studies of humans are performed by epidemiologists.

true association. Also called real association. The association that really exists between exposure to an agent and a disease and that might be found by a perfect (but nonetheless nonexistent) study.

Type I error. Rejecting the null hypothesis when it is true. See alpha error.

Type II error. Failing to reject the null hypothesis when it is false. See beta error.

validity. The degree to which a measurement measures what it purports to measure.

variable. Any attribute, condition, or other characteristic of subjects in a study that can have different numerical characteristics. In a study of the causes of heart disease, variables that might be measured are blood pressure and dietary fat intake.

variant. Any of multiple forms of a particular gene or region of DNA that is found in a population. Variants originally arise via mutations that change the DNA nucleotide base sequence. Variants may be harmful, beneficial, or may have no discernible effect. The functional significance of many variants is currently unknown. It is estimated that each person's genome contains three to four million variants.

References

On Epidemiology

- Anders Ahlbom & Steffan Norell, *Introduction to Modern Epidemiology* (2d ed. 1990).
- Casarett & Doull's *Toxicology: The Basic Science of Poisons* (Curtis D. Klaassen ed., 9th ed. 2019).
- Causal Inferences (Kenneth J. Rothman ed., 1988).
- David D. Celentano & Moyses Szklo, *Gordis Epidemiology* (6th ed. 2018).
- William G. Cochran, *Sampling Techniques* (1977).
- A Dictionary of Epidemiology (Miquel M. Porta et al. eds., 6th ed. 2014).
- Robert C. Elston & William D. Johnson, *Basic Biostatistics for Geneticists and Epidemiologists* (2008).
- Encyclopedia of Epidemiology (Sarah E. Boslaugh ed., 2008).
- Joseph L. Fleiss et al., *Statistical Methods for Rates and Proportions* (3d ed. 2003).
- Morton Hunt, *How Science Takes Stock: The Story of Meta-Analysis* (1997).
- International Agency for Research on Cancer (IARC), *Interpretation of Negative Epidemiologic Evidence for Carcinogenicity* (N.J. Wald & R. Doll eds., 1985).
- Harold A. Kahn & Christopher T. Sempes, *Statistical Methods in Epidemiology* (1989).
- Timothy L. Lash et al., *Modern Epidemiology* (4th ed. 2021).
- David E. Lilienfeld, *Overview of Epidemiology*, 3 Shepard's Expert & Sci. Evidence Q. 25 (1995).
- Marcello Pagano & Kimberlee Gauvreau, *Principles of Biostatistics* (3d ed. 2022).
- Richard K. Riegelman & Robert A. Hirsch, *Studying a Study and Testing a Test: How to Read the Health Science Literature* (5th ed. 2005).
- Bernard Rosner, *Fundamentals of Biostatistics* (7th ed. 2010).
- David A. Savitz, *Interpreting Epidemiologic Evidence: Strategies for Study Design and Analysis* (2003).
- James J. Schlesselman, *Case-Control Studies: Design, Conduct, Analysis* (1982).
- Dona Schneider & David E. Lilienfeld, *Lilienfeld's Foundations of Epidemiology* (4th ed. 2015).
- Brian L. Strom et al., *Pharmacoepidemiology* (6th ed. 2020).
- Lisa M. Sullivan, *Essentials of Biostatistics* (2008).
- Mervyn Susser, *Epidemiology, Health, & Society: Selected Papers* (1987).

On Law and Epidemiology

- American Law Institute, *Reporters' Study on Enterprise Responsibility for Personal Injury* (1991).

- Bert Black & David H. Hollander, Jr., *Unraveling Causation: Back to the Basics*, 3 U. Balt. J. Env't L. 1 (1993).
- Bert Black & David Lilienfeld, *Epidemiologic Proof in Toxic Tort Litigation*, 52 Fordham L. Rev. 732 (1984).
- Gerald Boston, *A Mass-Exposure Model of Toxic Causation: The Content of Scientific Proof and the Regulatory Experience*, 18 Colum. J. Env't L. 181 (1993).
- Vincent M. Brannigan et al., *Risk, Statistical Inference, and the Law of Evidence: The Use of Epidemiological Data in Toxic Tort Cases*, 12 Risk Analysis 343 (1992).
- Troyen Brennan, *Causal Chains and Statistical Links: The Role of Scientific Uncertainty in Hazardous-Substance Litigation*, 73 Cornell L. Rev. 469 (1988).
- Troyen Brennan, *Helping Courts with Toxic Torts: Some Proposals Regarding Alternative Methods for Presenting and Assessing Scientific Evidence in Common Law Courts*, 51 U. Pitt. L. Rev. 1 (1989).
- Philip Cole, *Causality in Epidemiology, Health Policy, and Law*, 27 Env't L. Rep. 10,279 (June 1997).
- Comment, *Epidemiologic Proof of Probability: Implementing the Proportional Recovery Approach in Toxic Exposure Torts*, 89 Dick. L. Rev. 233 (1984).
- George W. Conk, *Against the Odds: Proving Causation of Disease with Epidemiological Evidence*, 3 Shepard's Expert & Sci. Evidence Q. 85 (1995).
- Carl F. Cranor, *Toxic Torts: Science, Law, and the Possibility of Justice* (2006).
- Carl F. Cranor et al., *Judicial Boundary Drawing and the Need for Context-Sensitive Science in Toxic Torts After Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 16 Va. Env't L.J. 1 (1996).
- Richard Delgado, *Beyond Sindell: Relaxation of Cause-in-Fact Rules for Indeterminate Plaintiffs*, 70 Cal. L. Rev. 881 (1982).
- Dan B. Dobbs et al., *Dobbs' Law of Torts* (updated 2022).
- Michael Dore, *A Commentary on the Use of Epidemiological Evidence in Demonstrating Cause-in-Fact*, 7 Harv. Env't L. Rev. 429 (1983).
- Jean Macchiaroli Eggen, *Toxic Torts, Causation, and Scientific Evidence After Daubert*, 55 U. Pitt. L. Rev. 889 (1994).
- Daniel A. Farber, *Toxic Causation*, 71 Minn. L. Rev. 1219 (1987).
- Heidi Li Feldman, *Science and Uncertainty in Mass Exposure Litigation*, 74 Tex. L. Rev. 1 (1995).
- Stephen E. Fienberg et al., *Understanding and Evaluating Statistical Evidence in Litigation*, 36 Jurimetrics J. 1 (1995).
- Joseph L. Gastwirth, *Statistical Reasoning in Law and Public Policy* (1988).
- Herman J. Gibb, *Epidemiology and Cancer Risk Assessment*, in *Fundamentals of Risk Analysis and Risk Management* 23 (Vlasta Molak ed., 1997).
- Steve C. Gold, *The More We Know, The Less Intelligent We Are?—How Genomic Information Should, and Should Not, Change Toxic Tort Causation Doctrine*, 34 Harv. Env't L. Rev. 370 (2010).

Reference Guide on Epidemiology

- Steve Gold, *Causation in Toxic Torts: Burdens of Proof, Standards of Persuasion, and Statistical Evidence*, 96 Yale L.J. 376 (1986).
- Leon Gordis, *Epidemiologic Approaches for Studying Human Disease in Relation to Hazardous Waste Disposal Sites*, 25 Hous. L. Rev. 837 (1988).
- Michael D. Green, *Expert Witnesses and Sufficiency of Evidence in Toxic Substances Litigation: The Legacy of Agent Orange and Bendectin Litigation*, 86 Nw. U. L. Rev. 643 (1992).
- Michael D. Green, *The Future of Proportional Liability*, in *Exploring Tort Law* (Stuart Madden ed., 2005).
- Sander Greenland, *The Need for Critical Appraisal of Expert Witnesses in Epidemiology and Statistics*, 39 Wake Forest L. Rev. 291 (2004).
- Khristine L. Hall & Ellen Silbergeld, *Reappraising Epidemiology: A Response to Mr. Dore*, 7 Harv. Env't L. Rev. 441 (1983).
- Jay P. Kesan, *Drug Development: Who Knows Where the Time Goes?: A Critical Examination of the Post-Daubert Scientific Evidence Landscape*, 52 Food Drug Cosm. L. J. 225 (1997).
- Jay P. Kesan, *An Autopsy of Scientific Evidence in a Post-Daubert World*, 84 Geo. L. Rev. 1985 (1996).
- Constantine Kokkoris, Comment, *DeLuca v. Merrell Dow Pharmaceuticals, Inc.: Statistical Significance and the Novel Scientific Technique*, 58 Brooklyn L. Rev. 219 (1992).
- James P. Leape, *Quantitative Risk Assessment in Regulation of Environmental Carcinogens*, 4 Harv. Env't L. Rev. 86 (1980).
- David E. Lilienfeld, *Overview of Epidemiology*, 3 Shepard's Expert & Sci. Evidence Q. 23 (1995).
- Junius McElveen, Jr., & Pamela Eddy, *Cancer and Toxic Substances: The Problem of Causation and the Use of Epidemiology*, 33 Clev. St. L. Rev. 29 (1984).
- Modern Scientific Evidence: The Law and Science of Expert Testimony (David L. Faigman et al. eds., 2021–2022).
- Note, *Development in the Law—Confronting the New Challenges of Scientific Evidence*, 108 Harv. L. Rev. 1481 (1995).
- Susan R. Poulter, *Science and Toxic Torts: Is There a Rational Solution to the Problem of Causation?*, 7 High Tech. L.J. 189 (1992).
- Jon Todd Powell, Comment, *How to Tell the Truth with Statistics: A New Statistical Approach to Analyzing the Data in the Aftermath of Daubert v. Merrell Dow Pharmaceuticals*, 31 Hous. L. Rev. 1241 (1994).
- Restatement (Third) of Torts: Liability for Physical and Emotional Harm §§ 26, 28, cmt. c & rptrs. note (2010).
- David Rosenberg, *The Causal Connection in Mass Exposure Cases: A Public Law Vision of the Tort System*, 97 Harv. L. Rev. 849 (1984).
- Joseph Sanders, *The Bendectin Litigation: A Case Study in the Life-Cycle of Mass Torts*, 43 Hastings L. J. 301 (1992).

Reference Manual on Scientific Evidence

- Joseph Sanders, *Scientific Validity, Admissibility, and Mass Torts After Daubert*, 78 Minn. L. Rev. 1387 (1994).
- Joseph Sanders & Julie Machal-Fulks, *The Admissibility of Differential Diagnosis to Prove Causation in Toxic Tort Cases: The Interplay of Adjective and Substantive Law*, 64 L. & Contemp. Probs. 107 (2001).
- Palma J. Strand, *The Inapplicability of Traditional Tort Analysis to Environmental Risks: The Example of Toxic Waste Pollution Victim Compensation*, 35 Stan. L. Rev. 575 (1983).
- Richard W. Wright, *Causation in Tort Law*, 73 Cal. L. Rev. 1735 (1985).

Reference Guide on Toxicology

DAVID L. EATON, BERNARD D. GOLDSTEIN, AND
MARY SUE HENIFIN

David L. Eaton, Ph.D., is Professor Emeritus of Environmental and Occupational Health Sciences, and Former Dean and Vice Provost of the Graduate School at the University of Washington.

Bernard D. Goldstein, M.D., is Professor Emeritus of Environmental and Occupational Health, and Former Dean, School of Public Health at the University of Pittsburgh.

Mary Sue Henifin, J.D., M.P.H., is Of Counsel at Buchanan Ingersoll & Rooney, P.C.

CONTENTS

Introduction, 1030

Fundamental Principles of Toxicology, 1030

Safety and Risk Assessment, 1031

The Use of Toxicological Information in Risk Assessment, 1034

Toxicology and the Law, 1037

Toxicology and Epidemiology, 1038

The Role of Toxicology in Establishing Causation, 1039

Toxicology and Exposure Science, 1045

Use of Biomarkers in Toxicology Exposure Assessment, 1047

Biomarkers of Exposure, 1047

Biomarkers of Effect, 1048

Biomarkers of Susceptibility, 1049

Toxicological Study Design, 1049

In Vivo Research: Use of Live Animals in Toxicity Testing and
Safety Assessment, 1051

Determining Dose–Response Relationships, 1051

Acute Toxicity Testing and the Lethal Dose 50% (LD₅₀), 1052

No Observed Adverse Effect Level (NOAEL), 1052

Benchmark Dose (BMD), 1053

Linearized Non-Threshold (LNT) Model and Determination of
Cancer Risk, 1056

Maximum Tolerated Dose (MTD) and Chronic Toxicity Tests, 1058

In Vitro Research, 1059

New Approach Methodologies (NAMs), 1060

Extrapolation from Animal (In Vivo) and Cell (In Vitro) Research
to Humans, 1061

Toxicological Processes and Target Organ Toxicity, 1063

Demonstrating an Association Between Exposure and Risk of Disease, 1069

- On What Species of Animals Was the Compound Tested? What Is Known About the Biological Similarities and Differences Between the Test Animals and Humans? How Do These Similarities and Differences Affect the Extrapolation from Animal Data in Assessing the Risk to Humans?, 1069
- Does Research Show That the Compound Affects a Specific Target Organ? Will Humans Be Affected Similarly?, 1071
- What Is Known About the Chemical Structure of the Compound and Its Relationship to Toxicity?, 1072
- Has the Compound Been the Subject of In Vitro Research, and If So, Can the Findings Be Related to What Occurs In Vivo?, 1073
- Is the Association Between Exposure and Disease Biologically Plausible?, 1074
- Specific Causal Association Between an Individual's Exposure and Onset of Disease, 1074
 - Was the Plaintiff Exposed to the Substance, and If So, Did the Exposure Occur in a Manner That Can Result in Absorption into the Body?, 1076
 - Were Other Factors Present That Can Affect the Distribution of the Compound Within the Body?, 1078
 - What Is Known About How Metabolism in the Human Body Alters the Toxic Effects of the Compound?, 1078
 - What Excretory Route Does the Compound Take, and How Does This Affect Its Toxicity?, 1078
 - Does the Temporal Relationship Between Exposure and the Onset of Disease Support or Contradict Causation?, 1079
 - When Exposure to the Substance Is Associated with the Disease, Is There a Benchmark Dose or No Observed Adverse Effect Level, and If So, Was the Individual Exposed Above This Level?, 1080
 - What Is the Likelihood That the Disease Would Have Occurred Without the Specific Exposure?, 1081
- Medical History, 1085
 - Is the Medical History of the Individual Consistent with the Toxicologist's Expert Opinion Concerning the Injury?, 1085
 - Are the Complaints Specific or Nonspecific?, 1086
 - Do Laboratory Tests Indicate Exposure to the Compound?, 1087
 - What Other Causes Could Lead to the Given Complaint?, 1088
 - Is There Evidence of Interaction with Other Chemicals?, 1088
 - Do Humans Differ in the Extent of Susceptibility to the Particular Compound in Question? Are These Differences Relevant in This Case?, 1090
 - Has the Expert Considered Data That Contradict Their Opinion?, 1091
- Expert Qualifications, 1091
 - Does the Proposed Expert Have an Advanced Degree in Toxicology, Pharmacology, or a Related Field? If the Expert Is a Physician,

Reference Guide on Toxicology

Are They Board Certified in Toxicology or in a Field Such as Occupational Medicine?, 1092
Has the Proposed Expert Been Certified by the American Board of Toxicology, or Do They Belong to a Professional Organization, Such as the Academy of Toxicological Sciences or the Society of Toxicology?, 1094
What Other Criteria Does the Proposed Expert Meet?, 1095
Acknowledgments, 1096
Glossary of Terms, 1097
General References on Toxicology and Chemical Risk Assessment, 1104

FIGURES

1. Classical dose–response relationship showing various terms associated with points of departure for use in risk assessment, 1054
2. Idealized comparison of a linearized non-threshold and threshold dose–response relationship, 1057

TABLE

1. Sample of Selected Toxicological Endpoints and Examples of Agents of Concern in Humans, 1064

Introduction

The discipline of toxicology is concerned primarily with identifying and understanding the adverse effects of external chemical and physical agents on biological systems, including the prevention and amelioration of such adverse effects.¹ The field of risk assessment, when applied to chemicals, uses the basic principles of toxicology to help make informed decisions about how likely it is that a given proportion (e.g., 1 in 100,000) of individuals in an exposed population are likely to develop the toxic endpoint of concern. Such judgments are an essential component of regulatory toxicology in establishing health protective guidance and regulation for exposures to chemicals in the workplace, and chemical contaminants in air, drinking water, food, cosmetics, and pharmaceuticals.² In the context of toxic tort litigation, toxicology expert opinion is focused on the likelihood that an alleged exposure to an individual, or in some instances a group of individuals, was *more likely than not* a significant cause or contributor to the specific disease in question. Although it is an age-old science, toxicology has become a discipline distinct from pharmacology, biochemistry, cell biology, and related fields only in the past sixty years.

Fundamental Principles of Toxicology

There are three central tenets of toxicology. First, “the dose makes the poison”: This implies that all chemical agents are intrinsically hazardous; whether they cause harm is only a question of dose.³ Even water, if consumed in large quantities, can be toxic. Second, each chemical or physical agent tends to produce a specific pattern of biological effects that can be used to establish disease causation.⁴ Third, the toxic responses in laboratory animals are useful predictors of toxic responses in humans, with appropriate qualifications based on an understanding of how the chemical is metabolized in humans and experimental animals, and how the chemical causes its toxicity (so-called mechanism or mode of action). Each of these tenets, and their exceptions, is discussed in greater detail in this reference guide.

1. Casarett & Doull’s *Toxicology: The Basic Science of Poisons* (Curtis D. Klaassen ed., 9th ed. 2019) [hereinafter Casarett & Doull’s].

2. Elaine M. Faustman, *Risk Assessment*, in Casarett & Doull’s, *supra* note 1, at ch. 4.

3. For a discussion of more modern formulations of this principle, which was articulated by Paracelsus in the sixteenth century, see David L. Eaton, *Scientific Judgment and Toxic Torts—A Primer in Toxicology for Judges and Lawyers*, 12 J.L. & Pol’y 5, 15 (2003); Lauren M. Aleksunes & David L. Eaton, *Principles of Toxicology*, in Casarett & Doull’s, *supra* note 1, at ch. 2. A short review of the field of toxicology can be found in Curtis D. Klaassen, *Principles of Toxicology and Treatment of Poisoning*, in Goodman and Gilman’s *The Pharmacological Basis of Therapeutics* 1739 (11th ed. 2008).

4. Some substances, such as central nervous system toxicants, can produce complex and non-specific symptoms, such as headaches, nausea, and fatigue.

Reference Guide on Toxicology

The science of toxicology attempts to determine at what doses foreign agents produce their effects. The foreign agents classically of interest to toxicologists are all chemicals (including foods and pharmaceuticals) and physical agents in the form of radiation, but not living organisms that cause infectious diseases.⁵ However, the poisons that are produced by living organisms (properly called toxins), such as snake venom or botulinum toxin, are generally considered under the framework of toxicology. The discipline of toxicology provides scientific information relevant to the following questions:

1. What hazards does a chemical or physical agent present to human populations or the environment (ecotoxicology)?
2. What degree of risk is associated with chemical exposure at any given dose?⁶

This reference guide focuses on toxicology concepts used in regulatory, administrative, and toxic tort cases, with an emphasis on human health effects, rather than ecological effects. A set of model questions for evaluating the admissibility and strength of an expert's opinion is also included. Following each question is an explanation of the type of toxicological data or information that may be offered in response to the question, as well as a discussion of its significance. The relationship between toxicology, epidemiology, and exposure assessment is also addressed, with cross references to information included in the *Reference Guide on Epidemiology* and the *Reference Guide on Exposure Science and Exposure Assessment*.

Safety and Risk Assessment

Toxicological expert opinion also relies on formal safety and risk assessments. Safety assessment is the area of toxicology relating to the testing of chemicals and drugs for toxicity. It is a relatively formal approach in which the potential for toxicity of a chemical is tested in vivo or in vitro using standardized techniques. The protocols for such studies usually are developed through scientific consensus and are subject to oversight by governmental regulators or other watchdog groups.

After a number of bad experiences, including outright fraud, government agencies have imposed codes on laboratories involved in safety assessment,

5. Forensic toxicology, a subset of toxicology generally concerned with criminal matters, is not addressed in this reference guide, because it is a highly specialized field with its own literature and methodologies that do not relate directly to toxic tort or regulatory issues.

6. In standard risk assessment terminology, hazard—the ability of a chemical to cause (a) specific type(s) of toxic effect(s)—is an intrinsic property of a chemical or physical agent, while risk—the likelihood/probability that the effect will occur in an individual or population under a given set of circumstances—is dependent both upon hazard and on the extent of exposure (both the concentration or amount and the duration/frequency of exposure).

including industrial, contract, and in-house laboratories.⁷ Known as good laboratory practices (GLPs), these codes govern many aspects of laboratory standards, including such details as the number of animals per cage, dose and chemical verification, and the handling of tissue specimens. GLPs are remarkably similar across agencies, but the tests called for differ depending on the mission. For example, there are major differences between the Food and Drug Administration's (FDA) and the Environmental Protection Agency's (EPA) required procedures for testing drugs and environmental chemicals.⁸ The FDA requires and specifies both efficacy and safety testing of drugs in humans and animals. Carefully controlled clinical trials using doses within the expected therapeutic range are required for premarket testing of drugs, because exposures to prescription drugs are carefully controlled and should not exceed specified ranges or uses.

Although new chemicals manufactured for use as food additives, pharmaceuticals, pesticides, and cosmetics have required substantial toxicity testing prior to regulatory approval since at least the 1950s, the vast majority of industrial chemicals had little, if any, premarket toxicity testing until the mid-1970s. In 1976, Congress established new legislation, called the Toxic Substances Control Act (TSCA), to require chemical manufacturers of new chemical entities (chemicals not on an established list of the then approximately 60,000 chemicals in commerce) to submit premarket notifications that would require EPA evaluation and approval prior to the manufacture of significant quantities of the chemical. Although well intended, the legislation "lacked teeth" and over the years was widely viewed as ineffectual. In 2016, the TSCA was extensively revised under the Frank R. Lautenberg Chemical Safety for the 21st Century Act.⁹ This revised TSCA legislation (the Lautenberg Act) increases substantially the expectations for toxicity evaluation for existing chemicals, with a focus on those that the EPA believes pose the highest risk to human health and the environment, as well as giving the EPA more oversight prior to approval for the manufacturing of new chemicals. The EPA now provides extensive guidance to industry on the types of information and data

7. A dramatic case of fraud involving a pharmaceutical company reporting toxicology research to support the safety of a consumer product is described in *Schueneman v. Arena Pharmaceuticals, Inc.*, 840 F.3d 698 (9th Cir. 2016). The defendants made positive public statements to investors about the safety of its weight loss drug, specifically indicating that the drug was not carcinogenic and referred to supporting "animal studies." *Id.* at 700. However, it was discovered that the defendants were conducting a study in which the drug was given to lab rats, and the drug was in fact causing cancer in the rats. *Id.* at 701. The Ninth Circuit found that the plaintiff, one of the defendant's investors, successfully showed that the defendants intentionally and deliberately did not disclose the negative findings of the rat study to investors. *Id.* at 707–08.

8. *Comparison of FDA, EPA, OECD GLP* (Apr. 7, 2015), <https://www.fda.gov/inspections-compliance-enforcement-and-criminal-investigations/fda-bioresearch-monitoring-information/articles>.

9. See Frank R. Lautenberg Chemical Safety for the 21st Century Act, Pub. L. 114–182, 130 Stat. 448 (2016), <https://perma.cc/B4T2-9PV7>.

required under the TSCA Premanufacture Notification guidelines.¹⁰ While the law and regulatory guidance often do not explicitly specify which of the numerous available toxicity testing procedures must be performed, they usually do provide a framework. Such frameworks, which may vary between different regulations, enable both industry and the EPA to assess the potential harm a new chemical may pose to people or the environment.¹¹ Much of the evaluation relies upon chemical structure and comparison with other similar chemicals for which substantial toxicological and environmental data may be available.

In a similar manner, the European Union Regulation on Registration, Evaluation, Authorization and Restriction of Chemicals (REACH) requires extensive testing of both new chemicals and chemicals already in commerce.¹² Moreover, because exposures are less predictable, doses usually are given in a wider range in animal tests for nonpharmaceutical agents.¹³

10. See EPA, Points to Consider When Preparing TSCA New Chemical Notifications (June 2018), <https://perma.cc/H9G5-7SAT>.

11. The 2016 revisions of the TSCA set forth five possible EPA determinations on new chemical notices:

- The chemical or significant new use presents an unreasonable risk of injury to health or the environment;
- Available information is insufficient to allow the EPA to make a reasoned evaluation of the health and environmental effects associated with the chemical or significant new use;
- In the absence of sufficient information, the chemical or significant new use may present an unreasonable risk of injury to health or the environment;
- The chemical is or will be produced in substantial quantities and either enters or may enter the environment in substantial quantities or there is or may be significant or substantial exposure to the chemical; or
- The chemical or significant new use is not likely to present an unreasonable risk of injury to health or the environment.¹⁴

The lack of toxicity data for new chemicals (or new uses of existing chemicals) introduced into commerce has led the EPA to propose methods of evaluation using (1) in vitro toxicity pathway testing, (2) computational toxicology approaches using structural analogs for which toxicity information is available, (3) followed by whole-animal testing where warranted. See *EPA's Review Process for New Chemicals*, <https://perma.cc/U65D-WANT>.

12. REACH is a regulation of the European Union, adopted in 2007 to “improve the protection of human health and the environment from the risks that can be posed by chemicals, while enhancing the competitiveness of the EU chemicals industry. It also promotes alternative methods for the hazard assessment of substances in order to reduce the number of tests on animals.” See European Chemicals Agency, *Understanding REACH*, <https://perma.cc/AC3M-UTYF>.

The legislation was updated in 2022. See European Chemicals Agency, *Upcoming Changes to REACH Information Requirements*, <https://perma.cc/A6WV-L4GZ>.

13. The development of a new drug inherently requires searching for an agent that at useful doses has a biological effect (e.g., decreasing blood pressure), whereas those developing a new chemical for consumer use (e.g., a house paint) hope that at usual doses no biological effects will occur. There are other compounds, such as pesticides and antibacterial agents, for which a biological effect is desired, but it is intended that at usual doses humans will not be affected. These different expectations

A major impetus for both REACH and the Lautenberg Act was the recognition that less than 1% of the 60,000 to 75,000 chemicals in commerce had been subjected to a full safety assessment, and there were significant toxicological data on only 10% to 20% of them. Under the current U.S. and international approaches to testing chemicals with high production volume, the extent of toxicological information is expanding rapidly.¹⁵

Risk assessment is an approach increasingly used by regulatory agencies to estimate and compare the risks of hazardous chemicals and to assign priority for avoiding their adverse effects. The National Academy of Sciences defines four components of risk assessment: hazard identification, dose–response estimation, exposure assessment, and risk characterization.¹⁶

The Use of Toxicological Information in Risk Assessment

Risk assessment, as practiced by government agencies involved in regulating exposure to environmental chemicals, is highly dependent upon the science of toxicology and on the information derived from toxicological studies. The EPA, FDA, Occupational Safety and Health Administration (OSHA), Consumer Product Safety Commission, and other international (e.g., the World Trade Organization), national, and state agencies use risk assessment as a means to protect workers or the

are part of the rationale for the differences in testing information available for assessing toxicological effects. Under FDA rules, approval of a new drug usually will require extensive animal and human testing, including a randomized double-blind clinical trial for efficacy and toxicity. In contrast, under the TSCA, the only requirement before a new chemical can be marketed is that a premanufacturing notice be filed with the EPA, including any toxicity data in the company's possession.

14. See EPA, Regulatory Determinations made under Section 5 of the Toxic Substances Control Act (TSCA), <https://perma.cc/6FNB-R74U>

15. For a comparison of European REACH legislation and the U.S. EPA's 2016 TSCA provisions, see Ágnes Botos, John D. Graham & Zoltán Illés, *Industrial Chemical Regulation in the European Union and the United States: a Comparison of REACH and the Amended TSCA*, 22 J. Risk Rsch. 1187, <https://doi.org/10.1080/13669877.2018.1454495>. For an overview of current trends in toxicity testing, see Ida Fischer, Catherine Milton & Heather Wallace, *Toxicity Testing Is Evolving!*, 9 Toxicology Rsch. 67 (2020), <https://doi.org/10.1093/toxres/tfaa011>; see also Daniel Krewski et al., *Toxicity Testing in the 21st Century: Progress in the Past Decade and Future Perspectives*, 94 Archives Toxicology 1 (2020), <https://doi.org/10.1007/s00204-019-02613-4>.

16. Nat'l Rsch. Council, Science and Decisions: Advancing Risk Assessment (2009). See also *In re Johnson & Johnson Talcum Powder Prods. Mktg., Sales Pracs. & Prods. Litig.*, 509 F. Supp. 3d 116 (D.N.J. 2020) (expert opinion that talcum powder may cause ovarian cancer based on risk assessment and analysis of Bradford Hill factors considerations is admissible). Recently, a National Academy of Sciences panel has discussed potential approaches using systematic review to improve the EPA's risk paradigm. See Nat'l Acads. of Scis., Eng'g & Med., *The Use of Systematic Review in EPA's Toxic Substances Control Act Risk Evaluations* (2021).

public from adverse effects.¹⁷ Acceptable risk levels—for example, 1 in 1,000 to 1 in 1,000,000—are usually well below what can be measured through epidemiological study. Inevitably, this means that risk assessment is usually based solely on toxicological data—or, if epidemiological findings of an adverse effect are observed, then toxicological reasoning must be used to extrapolate to the appropriate lower exposure or dose standard aimed at protecting the public.

The four-part paradigm of risk assessment is heavily based on toxicological precepts. First, hazard identification, which is roughly equivalent to the legal term general causation, reflects the toxicological rule of specificity of effects, and the second step, dose–response assessment, is based upon “the dose makes the poison” and is fundamental to specific causation. The hazard identification process often uses weight-of-evidence approaches in which the toxicological, mechanistic, and epidemiological data are rigorously assessed to form a judgment regarding the likelihood that the agent produces a specific effect. Second, establishing the appropriate dose–response curve, threshold, or “one-hit” model is an exercise in toxicological reasoning. Even for those chemicals known to be carcinogens, a threshold model is appropriate if the toxicological mechanism of action can be demonstrated to depend upon a threshold. The third step, exposure assessment, requires knowledge of specific toxicological dynamics; for example, the impact on the lung of an air pollutant varies by factors such as: inhalation rate per unit body mass, which is affected by exercise and by age; the size of a particle or the solubility of a gas, either of which will affect the depth of penetration into the more sensitive parts of the airways; the competence of the usual airway defense mechanisms, such as mucus flow and macrophage function; and the ability of the lung to metabolize the agent.¹⁸ The fourth and final step is the characterization of risk.¹⁹

Risk assessment is not an exact science. It should be viewed as a useful framework to organize and synthesize information and to provide estimates on which policy making can be based. In recent years, codification of the methodology used to assess risk has increased confidence that the process can be reasonably free of bias; however, significant controversy remains, particularly when actual data are limited and generally conservative default assumptions are used.

17. Pharmaceuticals intended for human use are an exception in that a tradeoff between desired and adverse effects may be acceptable, and human data are available prior to, and as a result of, the marketing of the agent.

18. Some toxic agents pass through the lung without producing any direct effects on this organ. For example, inhaled carbon monoxide produces its toxicity in essence by being treated by the body as if it is oxygen. Carbon monoxide readily combines with the oxygen–combining site of hemoglobin, the molecule in red blood cells that is responsible for transporting oxygen from the lung to the tissues. In this way, the effective transport and tissue utilization of oxygen is blocked.

19. See, e.g., *Nat. Res. Defense Council v. U.S. EPA*, 38 F.4th 34 (9th Cir. 2022) (according to the EPA’s Office of Research and Development, the EPA did not follow its 2005 Guidelines for Carcinogen Risk Assessment, which lay out four steps for conducting risk assessments of chemicals’ carcinogenic potential).

Although risk assessment information about a chemical, and the research data on which it is based, may be useful in setting reasonable boundaries regarding the likelihood of causation, the impetus for the development of risk assessment has been the regulatory process, which has different goals.²⁰ Because of their use of appropriately prudent assumptions in areas of uncertainty and their use of default assumptions when there are limited data, as well as legal requirements related to public health considerations, risk assessments often intentionally encompass the upper range of possible risks.²¹ An additional issue, particularly related to cancer risk, is that standards based on risk assessment often are set to avoid the risk caused by lifetime exposure at this upper-range estimate of risk. Exposure to levels exceeding this standard for a small fraction of a lifetime does not mean that the overall lifetime risk of regulatory concern has been exceeded.²² Given these issues, courts have taken different approaches to how information derived from risk assessments may be relied upon by experts.²³

20. See Nat'l Rsch. Council & Comm. on Improving Risk Analysis Approaches Used by the U.S. EPA, *Toward a Unified Approach to Dose–Response Assessment, in Science and Decisions: Advancing Risk Assessment* 127 (2009), <https://perma.cc/47RG-NKMG>. See also *Yates v. Ford Motor Co.*, No. 5:12-CV-752-FL, 2015 WL 2189774, at *17, *27 (E.D.N.C. May 11, 2015) (holding that an EPA pamphlet that includes regulatory risk assessment can be relied on by the plaintiff's experts only to “support the reliability of studies the experts themselves rely upon”).

21. It is also claimed that standard risk assessment will underestimate true risks, particularly for sensitive populations exposed to multiple stressors, an issue of particular pertinence to discussions of environmental justice. Comm. on Env't Just., Institute of Med., *Toward Environmental Justice: Research, Education, and Health Policy Needs* (1999), <https://doi.org/10.17226/6034>. In 2003, the EPA developed formal guidance for cumulative risk assessment, wherein it defined aggregate exposure as the “combined exposure of an individual (or defined population) to a specific agent or stressor via relevant routes, pathways, and sources.” EPA, *Framework for Cumulative Risk Assessment* (2003), <https://perma.cc/D7CT-Z2RT>. In 2022, the EPA released a draft update to this document, with an initial emphasis on identifying research needs. See EPA, *Cumulative Impacts: Recommendations for ORD Research* (2022), <https://perma.cc/B6SL-EHUT>.

22. A public health standard to protect against the lifetime risk of inhaling a known carcinogen will usually be based on lifetime exposure calculations of twenty-four hours a day, every day for seventy years. This is more than 25,000 days and 600,000 hours. Exceeding this standard for a few hours would presumably have little impact on cancer risk. In contrast, for a short-term standard set to avoid a threshold-based risk, exceeding the standard for this short time may make a major difference—for example, an asthma attack caused by being outdoors on a day that the ozone standard is exceeded. Regulatory statutes sometimes limit consideration of multiple pathways of exposure to the same chemical, potentially underestimating overall exposure and risk. See, e.g., Swati D. G. Rayasam et al., *Toxic Substances Control Act (TSCA) Implementation: How the Amended Law Has Failed to Protect Vulnerable Populations from Toxic Chemicals in the United States*, 56 *Env't Sci. & Tech.* 11969 (2022), <https://doi.org/10.1021/acs.est.2c02079>. However, the interpretation of how such evaluations under the revised TSCA law are done by the EPA has been criticized. See Anthony C. Tweedale, *Correspondence on “Toxic Substances Control Act (TSCA) Implementation: How the Amended Law has Failed to Protect Vulnerable Populations from Toxic Chemicals in the United States,”* 56 *Env't Sci. & Tech.* 16533 (2022), <https://doi.org/10.1021/acs.est.2c06032>.

23. See, e.g., *Hardeman v. Monsanto Co.*, 997 F.3d 941 (9th Cir. 2021) (discussion of appropriate use in an expert opinion of the IARC's classification of glyphosate as a “probable carcinogen”).

Toxicology and the Law

Toxicological studies, by themselves, rarely offer direct evidence that a disease in any one individual was caused by a chemical exposure.²⁴ However, toxicology can provide scientific information regarding the increased risk of contracting a disease at any given exposure or dose and help rule out other risk factors for the disease (see section titled “Toxicology and Epidemiology” below for discussion on disease causation). Toxicological evidence also contributes to the weight of evidence supporting causal inferences by explaining how a chemical causes a specific disease through describing metabolic, cellular, and other physiological effects of exposure. For certain chemicals that create sufficiently long-lasting changes to molecular structures suitable for testing, toxicology evidence may be offered to establish chemical-specific “fingerprints” that can strengthen the causal inference between a specific chemical exposure and the presence of a specific disease.

The interface of toxicological science with legal issues can be complex, in part reflecting the inherent challenges of bringing science into a courtroom, but also because of issues pertinent to toxicology.

Complexity in toxicology is derived primarily from three factors. The first is that chemicals often change within the body as they go through various routes to eventual elimination.²⁵ Thus absorption, distribution, metabolism, and excretion are central to understanding the toxicology of an agent.

The second is that human sensitivity to chemical and physical agents can vary greatly among individuals, often as a result of differences in absorption, distribution, metabolism, or excretion. Although toxic responses from a given chemical sometimes occur in multiple different organs in the body (e.g., liver, kidney, brain), toxic effects often occur selectively in specific tissues, which are then referred to as target organs. Both the disposition of the chemical (e.g., metabolism and excretion) and target organ sensitivity can be affected by a combination of genetic and environmental factors, including factors as mundane as drinking certain fruit juices.²⁶

24. There are exceptions—for example, when measurements of levels in the blood or other body constituents of the potentially offending agent are at a high enough level to be consistent with reasonably specific health impacts, such as in carbon monoxide poisoning and lead poisoning.

25. Direct-acting toxic agents are those whose toxicity is due to the parent chemical entering the body. A change in chemical structure through metabolism usually results in detoxification. Indirect-acting chemicals are those that must first be metabolized to a harmful intermediate for toxicity to occur. For an overview of metabolism in toxicology, see David L. Eaton & Julia Cui, *Biotransformation*, in *Patty's Toxicology* (James Klaunig ed., 7th ed. 2023), <https://doi.org/10.1002/0471125474.tox122>.

26. Grapefruit juices and certain other fruit juices at sufficient dose have the potential to inhibit the metabolism of a wide variety of drugs and chemicals. See Michael J. Hanley et al., *The Effect of Grapefruit Juice on Drug Disposition*, 7 *Expert Op. Drug Metabolism & Toxicology* 267 (2011), <https://doi.org/10.1517/17425255.2011.553189>; Meng Chen et al., *Food-Drug Interactions*

The third major source of complexity is the need for extrapolation—either across species, because many toxicological data are obtained from studies in laboratory animals, or across doses, because human toxicological and epidemiological data often are limited to specific exposure or dose ranges that may differ from the ranges of interest. All three of these factors are responsible for much of the complexity in utilizing toxicology for tort or regulatory judicial decisions and are described in more detail below.

The science of toxicology is useful in establishing general causation in toxic tort litigation through demonstrating that specific chemicals are capable of causing specific effects (e.g., specific types of cancers, birth defects, poisoning) and identifying when such effects may be manifested, either acutely or over time. Identifying such qualitative characteristics of a chemical is referred to as hazard evaluation and is often based on results of experimental animal studies. To determine the likelihood that a specific chemical caused a specific toxic effect in an individual person (specific causation in tort litigation) or a population (regulatory/population health guidance) requires careful consideration of the exposure characteristics (how much, when, and how often) that are central to determining the dose–response for the specific toxic endpoint in the individual or population. Such decisions are complex and may also rely on human epidemiological data where they exist.

Toxicology and Epidemiology

From the beginning of human history, “Why me?” is probably the most frequently asked question about disease. Epidemiology is the scientific discipline most closely involved in providing evidence to answer this question. Epidemiology is the study of the incidence and distribution of disease in human populations. Clearly, both epidemiology and toxicology have much to offer in elucidating the causal relationship between chemical exposure and disease.²⁷ These sciences often go hand in hand with assessments of the risks of chemical exposure, without

Precipitated by Fruit Juices Other than Grapefruit Juice: An Update Review, 26 J. Food & Drug Analysis S61 (2018), <https://doi.org/10.1016/j.jfda.2018.01.009>.

27. See Steve C. Gold et al., *Reference Guide on Epidemiology*, section titled “General Causation,” in this manual. See also Nat’l Acads. Sci., Eng’g & Med, *Assessing Causality from a Multidisciplinary Evidence Base for National Ambient Air Quality Standards* (2022), <https://perma.cc/6L9G-NBCN> (discussing the strengths and limitations of epidemiological studies in establishing “causation” between exposure to a hazardous chemical (in this case, various air pollutants) and the development of specific diseases). For example, in *In re Nexium Esomeprazole*, 662 F. App’x 528 (9th Cir. 2016), testimony on general causation was excluded as unreliable in which the expert did not adequately explain how he inferred a causal relationship from epidemiological studies that did not come to the same conclusions. However, epidemiological studies are not always necessary. *McMunn v. Babcock & Wilcox Power Generation Grp., Inc.*, No. 2:10CV143, 2014 WL 814878, at *15 (W.D. Pa. Feb. 27, 2014) (citing *Glastetter v. Novartis Pharms. Corp.*, 252 F.3d 986, 992 (8th Cir. 2001); *Rider v. Sandoz Pharms. Corp.*, 295 F.3d 1194, 1198 (11th Cir. 2002)).

artificial distinctions being drawn between them. However, although courts generally rule epidemiological expert opinion admissible, the admissibility of toxicological expert opinion has been more controversial because of uncertainties regarding extrapolation from animal and in vitro data to humans. This has been particularly true in cases in which relevant epidemiological research data exist. However, the methodological weaknesses of some epidemiological studies, including their inability to accurately measure exposure, the often small numbers of subjects, and issues such as recall bias when individuals with specific diseases are asked to recall their past exposure to chemicals in comparison with a control population, often render these studies difficult to interpret. As detailed in the *Reference Guide on Epidemiology*, an epidemiological study by itself can only show an association, which may or may not be causal.

The Role of Toxicology in Establishing Causation

The most well-known guidelines for considering causality of a specific effect by a specific cause are the Bradford Hill considerations.²⁸ Austin Bradford Hill was a British epidemiologist who contributed greatly to studies linking cigarette smoking to lung cancer. As discussed in detail in the *Reference Guide on Epidemiology*, Hill described nine points of consideration to facilitate the determination of a causal association seen in an epidemiology study. A modified list of the Bradford Hill considerations that inform epidemiologists in making judgments about causation includes:

1. Temporal relationship
2. Strength of the association
3. Dose–response relationship
4. Replication of the findings
5. Biological plausibility
6. Consideration of alternative explanations
7. Cessation of exposure
8. Specificity of the association
9. Consistency with other knowledge

28. Austin Bradford Hill, *The Environment and Disease: Association or Causation?*, 108 J. Royal Soc’y Med. 32 (2015) (reprint of 1965 article), <https://doi.org/10.1177/0141076814562718>. For a discussion of the Bradford Hill considerations, see Steve C. Gold et al., *Reference Guide on Epidemiology*, section titled “General Causation,” in this manual. Although the points raised by Bradford Hill in his 1965 article are often referred to as “criteria,” the author made it clear that none were absolutes. As discussed in the *Reference Guide on Epidemiology* in this manual, the term “consideration,” rather than “criteria” or even “guidelines,” is perhaps more appropriate to represent these ideas.

Toxicological evidence can provide insights into almost all of these considerations:

1. **Temporal relationship** is perhaps obvious and refers to a fundamental concept that, in order for a toxic substance to have caused an adverse effect in an individual (or population of individuals), the *exposure must have occurred before the onset of symptoms* of the disease. However, the temporality consideration would not exclude the potential for a chemical to enhance or aggravate a preexisting disease. Toxicology often considers the potential interaction of two chemicals, referred to as potentiation, where chemical A does not by itself cause a specific adverse response or disease, but when combined with exposure to chemical B, which has been established to cause the adverse response or disease in question, the magnitude of response to chemical B is much greater in the presence of chemical A. A classic example of potentiation would be someone who takes a statin drug to lower cholesterol, with no evident adverse effects, but, upon regular consumption of large amounts of grapefruit juice, develops muscle weakness and spasms (rhabdomyolysis), a symptom of statin toxicity. In this example, the causal agent is actually the statin drug, but effects occur only after consuming large amounts of grapefruit juice.
2. **Strength of association** refers to the overall magnitude of effect and is largely pertinent to epidemiology studies.
3. **Dose–response** is a fundamental tenet of toxicology and is highly relevant to both toxicological and epidemiological studies. This is discussed in greater detail in other parts of this reference guide.
4. **Replication of the findings** is also relevant to both epidemiology and toxicology. For toxicology, this is most important in demonstrating another Bradford Hill consideration, biological plausibility, whereby several independent studies under different conditions demonstrate that chemical A reproducibly causes effect B. This reproducibility could be in animal studies, in vitro studies with human tissues, or both.
5. **Biological plausibility** is, of all of the Bradford Hill considerations, perhaps the most relevant to toxicology. As discussed in detail later in this reference guide, understanding the cellular, biochemical, and molecular processes (mechanisms and mode of action) by which a toxic substance causes an adverse effect is a critical element that can support a causal association in epidemiology studies, largely by demonstrating that adverse effects seen in experimental animals are relevant to humans because they share common mechanisms/modes of action, supporting biological plausibility. Because of ethical considerations, with the exception of pharmaceutical agents, controlled in vivo studies are largely limited to experimental animals, which is the domain of toxicology.

6. **Consideration of alternative explanations** is largely confined to epidemiology, although the consideration of mechanisms/modes of action of a toxic substance determined in experimental animals (the domain of toxicology) may be informative of alternative explanations—for example, a workplace where exposure to multiple different substances is associated with an elevated incidence of disease. Determining which, if any, of the substances has been shown through toxicological studies to cause the disease in question may support an alternative explanation.
7. **Cessation of exposure** can also be supported by animal studies in which the toxic response or disease in question is reduced or eliminated following cessation of exposure. This would be similarly informative and provide biological plausibility for such observations seen in epidemiology studies.
8. **Specificity of the association** is also relevant to toxicology, where a specific adverse effect can be shown to cause an unusual or uncommon effect in animals. Toxicological studies on mechanism or mode of action may demonstrate a specific or even unique development of a particular disease or symptom in experimental animals and/or human tissues in vitro, lending biological plausibility to the specificities of association observed in epidemiology studies.
9. **Consistency with other knowledge** also fits equally well in toxicology studies, again focused on a scientific understanding of how the toxic substance in question causes a specific adverse outcome or disease.

Thus, toxicology may contribute to the Bradford Hill considerations, for consistency considerations, particularly biological plausibility. Making sense of whether an association is causal is affected by whether there is a likely biological pathway between the exposure and the effect, for which the evidence is often based on animal toxicology studies. However, Hill argues that biological plausibility cannot be demanded, because it depends upon the biological knowledge of the day. The importance of biological plausibility to setting environmental regulation is evident in that it is now being routinely included as a chapter or subchapter in EPA analyses on which air pollution and other standards are based.

In recent decades, medical science has been challenged by the need to derive actionable information from the medical literature—a breadth of literature that has increased both in size and in the diversity of medical specialties and subspecialties that are being applied to solve problems related to diagnosis and treatment of disease. This has led to the development of formal systematic approaches to gathering and evaluating the pertinent epidemiological literature. The National Institutes of Health (NIH) held consensus-developing conferences from 1977 to 2013 aimed at controversial topics in clinical decision-making, which included

toxicological data where relevant. These were discontinued as no longer necessary because many other organizations had developed similar approaches.²⁹

A more recent example of a common approach to developing a systematic review process for evaluating the medical science literature has been that of the Cochrane reviews.³⁰ These are systematic approaches that analyze the existing literature and focus almost exclusively on epidemiological findings, primarily related to diagnosis and treatment of disease. In part building on the Cochrane review model, systematic reviews have been suggested for toxicology³¹ and for public health.³²

In contrast to epidemiology, because animal and cell studies permit researchers to isolate the effects of exposure to a single chemical or to mixtures, toxicological findings offer unique information concerning dose–response relationships, mechanisms of action, specificity of response, and other information relevant to the assessment of causation—although these can present issues related to extrapolation to humans.³³

The gold standard in clinical epidemiology and in the testing of pharmaceutical agents is the randomized double-blind placebo control study in which the control and intervention groups are usually well matched. Although appropriate and very informative for the testing of pharmaceutical agents, it is generally unethical for chemicals used for other purposes. The randomized control design in essence is what is used in a classic toxicological study in laboratory animals,

29. For a review of consensus approaches in general and the importance of unbiased selection of members of review committees that reflect the broad disciplines involved, see Bernard D. Goldstein, *What the Trump Administration Taught Us About the Vulnerabilities of EPA's Science-Based Regulatory Processes: Changing the Consensus Processes of Science into the Confrontational Processes of Law*, 31 Health Matrix 299 (2021), <https://perma.cc/YJ3H-TPPR>. For a discussion of general issues regarding medical practice and testimony, see John B. Wong et al., *Reference Guide on Medical Testimony*, in this manual.

30. Cochrane Library, About Cochrane Reviews, <https://perma.cc/UXR9-4HPL>.

31. Lena Smirnova et al., *3S—Systematic, Systemic, and Systems Biology and Toxicology*, 35 ALTEX 139 (2018) <https://doi.org/10.14573/altex.1804051>. See also Thomas Hartung et al., *Systems Toxicology II: A Special Issue*, 30 Chem. Rsch. Toxicology 869 (2017), <https://doi.org/10.1021/acs.chemrestox.7b00038>.

32. See also Steve C. Gold et al., *Reference Guide on Epidemiology*, section titled “Methods for Synthesizing or Combining the Results of Multiple Studies,” in this manual.

33. Courts have found animal studies to be relevant evidence of causation where there is a sound basis for extrapolating conclusions from those studies to humans in real-world conditions. See *Hardeman v. Monsanto Co.*, 997 F.3d 941 (9th Cir. 2021); see also *Milana v. Eisai, Inc.*, No. 8:21-cv-831-CEH-AEP, 2022 WL 846933 (M.D. Fla. Mar. 22, 2022) (increase in tumors in mammary tissue of experimental animals compared to higher rates of cancer in patients taking weight loss drug); *In re Levaquin Prods. Liab. Litig.*, No. 08–1943 (JRT), 2010 WL 8400514, at *2–4, *8 (D. Minn. Nov. 8, 2010); *Brandeis Univ. v. Keebler Co.*, No. 1:12-cv–01508, 2013 WL 5911233 (N.D. Ill. Jan. 18, 2013); *In re Abilify (Aripiprazole) Prods. Liab. Litig.*, 299 F. Supp. 3d 1291, 1310 (N.D. Fla. 2018).

although matching is more readily achieved because the animals are bred to be genetically similar and have identical environmental histories.

Dose issues are at the intersection between toxicology and epidemiology. Many epidemiological studies of the potential risk of chemicals do not have direct information about dose, although qualitative differences among subgroups or in comparison with other studies can be inferred. The epidemiology database includes many studies that are probing for the potential for an association between a cause and an effect. Thus, a study asking all those suffering from a specific disease a multiplicity of questions related to potential exposures is bound to find some statistical association between the disease and one or more exposure conditions due to chance alone. Such studies generate hypotheses that can then be evaluated by subsequent studies that more narrowly focus on the potential cause-and-effect relationship. One way to evaluate the strength of the association is to assess whether those epidemiological studies evaluating cohorts with relatively high exposure support the association.³⁴

The requirement in certain jurisdictions for epidemiological evidence of a relative risk greater than two ($RR > 2$) for general causation also has limited the utilization of toxicological evidence.³⁵ A firm legal requirement for such evidence means that if the epidemiological database only showed statistically significant evidence that cohorts exposed to ten parts per million of an agent for twenty years produced an 80% increase in risk, the court could not hear the case of a plaintiff alleging that exposure to fifty parts per million for twenty years of the same agent caused the adverse outcome. Yet to a toxicologist, there would be little question that exposure to the fivefold-higher dose of an agent producing an 80% increase in risk would lead to more than a 100% increase in risk, i.e., more than doubling of the risk.

34. As an example of considering dose and exposure issues across epidemiological studies, see Luoping Zhang et al., *Formaldehyde Exposure and Leukemia: A New Meta-Analysis and Potential Mechanisms*, 681 *Mutation Resch.* 150 (2008), <https://doi.org/10.1016/j.mrrev.2008.07.002>. The subject of the strength of an epidemiological association and its relation to causality is considered in Steve C. Gold et al., *Reference Guide on Epidemiology*, in this manual; see also note 79, *infra*.

35. The basis for the use of $RR > 2$ is the translation of the preponderance of evidence rule, or “more likely than not” legal standard, into the epidemiological concept of at least a doubling of risk. An example is the *Havner* rule in Texas, which for general causation requires that there be reliable epidemiological studies with a statistically significant $RR > 2$ associating a putative cause with an effect. See *Bostic v. Ga. Pac. Corp.*, 439 S.W.3d 332 (Tex. 2014). But see *Hardeman v Monsanto Co.*, 997 F.3d 941, 966 (9th Cir. 2021) (“[W]e have never suggested that a headline increase in a risk statistic, or even an adjusted odds ratio above 2.0, is necessary for finding a strong association. To the contrary, flexibility is warranted considering the contextual nature of the *Daubert* inquiry.”) (citations omitted). For a discussion of the use by courts in various jurisdictions of relative risk > 2 for general and specific causation, see Russel S. Carruth & Bernard D. Goldstein, *Relative Risk Greater Than Two in Proof of Causation in Toxic Tort Litigation*, 41 *Jurimetrics* 195 (2001); for the toxicological issues, see Bernard D. Goldstein, *Toxic Torts: The Devil Is in the Dose*, 16 *J.L. & Pol’y* 551–85 (2008).

Even though there is little toxicological data on many of the approximately 75,000 compounds in general commerce, there is usually far more information from toxicological studies than from epidemiological studies.³⁶ It is much easier, and usually more economical, to expose an animal to a chemical or to perform *in vitro* studies than it is to perform epidemiological studies. This difference in data availability is evident even for cancer causation, for which toxicological study is particularly expensive and time consuming. Of the approximately fifty specific chemicals that reputable international authorities agree are known human carcinogens based on positive epidemiological studies, almost all of them have also been shown to cause cancer in laboratory animal studies. Yet there are hundreds of known animal carcinogens for which there is either no valid epidemiological database or for which the epidemiological database has been equivocal.³⁷

To clarify any findings, regulators can require a repeat of an equivocal two-year animal toxicological study or the performance of additional laboratory

36. See, for example, Nat'l Toxicology Prog., 15th Rep. on Carcinogens (2021), <https://perma.cc/V4DP-3YNZ>, for a discussion of how epidemiology and animal toxicity studies are used to assess the potential cancer risk of different chemicals.

37. See, e.g., Int'l Agency for Rsch. on Cancer, IARC Monographs on the Identification of Carcinogenic Hazards to Humans, <https://perma.cc/R56M-EQMM>; Nat'l Toxicology Prog., *supra* note 36; *Waite v. All Acquisition Corp.*, 194 F. Supp. 3d 1298 (S.D. Fla. 2016) (court remains "unconvinced" that lack of statistically significant epidemiological studies properly overcomes all other types of scientific evidence including animal, cellular, and molecular studies). The absence of epidemiological data is due, in part, to the difficulties in conducting cancer epidemiology studies, including the lack of suitably large groups of individuals exposed for a sufficient period of time, long latency periods between exposure and manifestation of disease, the high variability in the background incidence of many cancers in the general population, and the inability to measure actual exposure levels. See also *Parker v. Mobil Oil Corp.*, 857 N.E.2d 1114 (N.Y. 2006), in which New York's highest court reviewed an appeal from a summary judgment of a lower court that had dismissed a plaintiff's case because of the lack of a specific exposure estimate by the plaintiff's experts. The plaintiff's experts claimed that unusual work practices during his seventeen years employed at a gasoline station led to higher exposure to benzene, a known cause of acute myelogenous leukemia (AML) from which the patient suffered, than would have occurred in the work practices of a usual gasoline station worker. The higher court found for the plaintiff on the lack of a need for a specific exposure assessment, citing other cases finding that an expert does not need to establish the amount of exposure the plaintiff experienced. *Id.* at 1121 (citing *McClain v. Metabolife Int'l, Inc.*, 401 F.3d 1233, 1241 n.6 (11th Cir. 2005)). However, the same court ruled for the defense based on the absence of epidemiological evidence that gasoline caused AML, despite the fact that benzene is a component of the gasoline mixture. Court preference for epidemiology is also evident in the *Havner* rule (see *supra* note 35). The requirement for epidemiological evidence of a doubling of risk precludes toxicological reasoning related to dose. For example, suppose there were two epidemiological studies showing a statistically significant 80% increase of an adverse endpoint in large workforces in well-regulated industries that were exposed to ten parts per million (ppm) of a chemical for thirty years. Suppose also that there was a plaintiff with the same endpoint whose experts presented evidence that the plaintiff was exposed to fifty ppm for thirty years to the same chemical. Based on dose considerations, a toxicologist could argue that the plaintiff's level of exposure would exceed 100%, i.e., more than double the plaintiff's risk. But under a strict interpretation of the *Havner* rule, the plaintiff's case would be dismissed.

studies in which animals deliberately are exposed to the chemical. Such deliberate exposure is not possible in humans. As a general rule, unequivocally positive epidemiological studies reflect prior workplace practices that led to relatively high levels of chemical exposure for a limited number of individuals and that, fortunately, in most cases no longer occur now. Thus, an additional prospective epidemiological study often is not possible, and even the ability to do retrospective studies is constrained by the passage of time.

In essence, epidemiological findings consistently demonstrating an adverse effect in humans of a manufactured chemical represent a failure of toxicology as a preventive science. A corollary of the tenet that, depending upon dose, all chemical and physical agents are harmful is that society depends upon toxicological science to discover and anticipate these harmful effects, and we also rely on regulators and other responsible parties to prevent human exposure to a harmful level or to ensure that the agent is not manufactured and distributed. In this regard epidemiology is a valuable backup approach that functions to detect failures of primary prevention. The two disciplines complement each other, particularly when the approaches are iterative.

Toxicologists may use epidemiological findings in their expert opinions concerning general causation and specific causation. This is particularly true for well-studied compounds that have been thoroughly evaluated by respected organizations, such as the International Agency for Research on Cancer (IARC, affiliated with the World Health Organization (WHO)) or the National Toxicology Program (NTP, affiliated with the U.S. Department of Health and Human Services), which include toxicologists and epidemiologists in their review of the evidence. For example, in mortality studies of large cohorts of workers exposed to high levels of a carcinogenic agent in which all types of cancer are reported, but only an increase in brain cancer is noted, the toxicologist may consider this information in determining if a pancreatic cancer was caused by exposure to a lesser dose of the agent. In terms of assessing specific causation for exposure to this known carcinogen, there may be the equivalent of a No Observed Adverse Effect Level (NOAEL; see section titled “Toxicological Study Design” below) in the reviewed and accepted epidemiological data in which an increase in brain cancer risk is seen only in the more highly exposed populations. Whether a toxicologist is qualified to offer an opinion based on both toxicology and epidemiology data depends on the particular background of the proffered expert and the legal standard applied in the jurisdiction.

Toxicology and Exposure Science

In recent decades, exposure assessment has developed into a scientific field (exposure science) with journals, learned societies, and research funding processes.

Exposure science methodologies include mathematical models used to predict exposure resulting from an emission source, which might be a long distance upwind; chemical or physical measurements of media such as air, food, and water; and biological monitoring within humans, including measurements of blood and urine specimens if available, for the chemical of concern and/or its metabolites. Extrapolation back to the level of external exposure depends upon toxicokinetic analyses (see section titled “Safety and Risk Assessment” above). In this continuum of exposure metrics, the closer to the human body, the greater the overlap with toxicology.³⁸ An exposure assessment should also look for competing exposures from other sources of the chemical of concern.

Exposure assessment is central to epidemiology as well and is often the limiting factor in using an epidemiological study to infer causation. Many of the causal associations between chemicals and human disease have been developed from epidemiological studies relating a workplace chemical to an increased risk of the specific disease in cohorts of workers, often with only a qualitative assessment of exposure. More sophisticated modern approaches to chemical exposures at complex workplaces include development of a job exposure matrix based on specific exposure levels throughout the working lifetime. An improved quantitative understanding of such exposures enhances the likelihood of observing causal relations.³⁹ It also can support the opinion of the expert toxicologist on the likelihood that a specific exposure was responsible for an adverse outcome.

38. Toxicologists also have indirect means of approaching exposure through symptoms— e.g., for many agents, there is a known threshold for smell and a reasonable range of levels that might cause symptoms. For example, the use of toxicological expertise is appropriate in a situation in which chronic exposure to a volatile hydrocarbon is alleged to have occurred at levels at which acute exposure would be expected to render the individual unconscious. Toxicologists may also contribute knowledge of the extent of individual exposure based upon appropriate assumptions concerning inhalation rate or water use; for example, children inhale more per body mass than do adults, and outdoor workers in hot climates will drink more fluids than those in cool climates. For a discussion of general issues regarding assessment of exposure of toxic substances, see M. Elizabeth Marder & Joseph V. Rodricks, *Reference Guide on Exposure Science and Exposure Assessment*, in this manual.

39. In terms of general causation, accurate exposure assessment is important because a true effect can be missed owing to confounding caused by cohorts that often include workers with little exposure to the putative offending agent, thereby diluting the actual effect. See Peter F. Infante, *Benzene Exposure and Multiple Myeloma: A Detailed Meta-analysis of Benzene Cohort Studies*, 1076 *Annals N.Y. Acad. Scis.* 90 (2006), <https://doi.org/10.1196/annals.1371.081>, for a discussion of this issue in relation to a meta-analysis of the potential causative role of benzene in multiple myeloma. On the other hand, an association between exposure and effect occurring solely by chance is more likely if the effect does not meet the expected standard of being more pronounced in those receiving the highest dose. See Goldstein, *supra* note 35. Setting regulatory standards based on the observed effect in a cohort often requires a risk assessment, which in turn is dependent on understanding the extent of the exposure. This has led to extensive retrospective reconstruction of exposure in key cohorts.

Use of Biomarkers in Toxicology Exposure Assessment

The term biomarker is used to identify the measurement of a chemical or biological molecule present in tissue when such tests are available for the substance at issue, and where the tests have been administered within the time frame when the substance may be present in the body to indicate past and/or current exposure.⁴⁰ There are three different types of biomarkers of relevance in toxicology: a) biomarkers of exposure, b) biomarkers of effect, and c) biomarkers of susceptibility. Biomarkers may be persistent in tissue or may rapidly disappear, depending on the chemicals at issue.

Biomarkers of Exposure

A widely used biomarker of exposure is blood alcohol levels estimated by the Breathalyzer test, frequently used by law enforcement to establish whether an automobile driver is inebriated. The science behind the breathalyzer test has been the subject of many court decisions. In a similar manner, measurements of the concentration of drug(s) in the blood may be used to establish cause of death. However, alcohol and most drugs have relatively short half-lives (the time it takes for 50% of the chemical to disappear from the blood), on the order of a few hours or less, and thus are useful only if the blood (or breath) sample is collected soon after the exposure occurred.

The concentration of lead in the blood and/or bones of an individual is another example of a biomarker of exposure that has been extensively studied. In the case of lead exposure, the half-life in blood is about one month, so repeated exposures in the workplace or environment can result in blood lead concentrations above what is measured in the environment. The half-life of lead in bone is on the order of twenty-five to thirty years, so assessment of lead in bone, performed noninvasively through an X-ray fluorescence technique, provides a window into past exposures to lead over a lifetime. Some other environmentally relevant toxic substances, referred to collectively as persistent organic pollutants or POPs (e.g., polychlorinated biphenyls (PCBs), dioxins, certain organochlorine pesticides like DDT and chlordane, and more recently a group of fluorinated chemicals collectively called per- and polyfluoroalkyl substances, or PFAS⁴¹),

40. See Nat'l Rsch. Council, *Human Biomonitoring for Environmental Chemicals* (2006), <https://perma.cc/9AQ2-QXUQ>; see also Lesa L. Aylward, *Integration of Biomonitoring Data into Risk Assessment*, 9 *Current Op. Toxicology* 14 (2018), <https://doi.org/10.1016/j.cotox.2018.05.001>.

41. Sometimes also referred to in the popular press as “forever chemicals.” For an EPA summary of research on PFAS, see <https://perma.cc/JLV4-54XX>.

have half-lives measured in years, and thus slow accumulation from environmental exposures can occur over many years. For these types of chemicals, blood concentrations are indicative of lifetime exposures—not just recent exposure—and can be useful to establish whether an unusual exposure to a specific agent has occurred in the past.

The Centers for Disease Control and Prevention's (CDC) National Health and Nutrition Examination Survey (NHANES) monitors a broad level of biological markers of health—e.g., blood cholesterol levels—on a U.S. population sample. This survey now includes measurements of a wide variety of environmental and dietary substances. For environmental agents, the CDC considers blood levels above the statistical ninety-fifth percentile level in the NHANES samples as an indication of an “unusual exposure,” although it does not state whether exposures above, or below, the ninety-fifth percentile are necessarily harmful or safe, respectively.⁴² However, other regulatory agencies might use such levels to indicate that an exposure at that level is associated with an unacceptable risk.

Biomarkers of Effect

Research on certain toxic chemicals has established the presence of biological changes indicative of the effects of that specific chemical and, thus, can also provide evidence of exposure, at least within the time period that the biological changes remain within the body. For example, when carbon monoxide enters the bloodstream, it binds to hemoglobin, the molecule responsible for delivering oxygen from the lungs to the rest of the body. Once bound, it prevents oxygen from binding, which results in oxygen deprivation that can be rapidly fatal. For sublethal doses of carbon monoxide, the time to full equilibrium with blood hemoglobin is eight to twelve hours, depending largely on breathing rate. The measurement of the amount of this carboxyhemoglobin in blood is therefore both a biomarker of exposure—reflecting carbon monoxide exposure during the past eight to twelve hours—as well as a biomarker of effect, as it is on the direct pathway to toxicity. Carbon monoxide can also be measured in expired air, in which case it primarily reflects exposure. Another example of a biomarker of exposure and effect is provided by a broad class of insecticides known as organophosphorus chemicals (OPs). As a chemical class, OPs bind to and inhibit an important

42. For more details on the NHANES program, see <https://perma.cc/UY4L-S8TY>. One objective of the NHANES program was to “establish reference ranges that physicians and scientists can use to determine whether a person or group has an unusually high exposure. . . . The 95th percentile is helpful when determining whether levels observed in separate public health investigations or other studies are unusual.” CDC, Fourth National Report on Human Exposure to Environmental Chemicals (Volume Three: Analysis of Pooled Serum Samples for Select Chemicals, NHANES 2005–2016), <https://perma.cc/E28W-X5HL>.

enzyme in nerve function called cholinesterase and can be rapidly lethal. Cholinesterase activity unrelated to its role in the nervous system is also present in red blood cells and in plasma. Measurements of cholinesterase in the blood can thus be an effects-related biomarker of exposure to many OP pesticides. This test is used frequently in the occupational environment and clinically to diagnose OP poisoning. Where available, and when testing is done within the required time frames, such effects-related biomarkers provide information that is directly on the pathway between external exposure and internal effects. Although currently there are only a relatively few chemicals for which specific biomarkers are available, it is predicted that the increasing understanding of molecular effects of toxic substances will lead to more effects-related biomarkers of exposure.

Biomarkers of Susceptibility

Just as biological measurement of environmental exposures is an area of research interest, another area of research interest is DNA analyses to identify individuals who may be more susceptible to exposures, such as genetically determined variability in enzymes responsible for detoxifying chemicals of concern. Given the complexities of environmental and genetic interactions, much discussion in this area remains theoretical.

Toxicological Study Design

Toxicological studies usually involve exposing laboratory animals (in vivo research) or cells or tissues (in vitro research) to chemical or physical agents, monitoring the outcomes (such as cellular abnormalities, tissue damage, organ toxicity, or tumor formation), and comparing the outcomes with those for unexposed control groups. As explained below,⁴³ the extent to which animal and cell experiments accurately predict human responses to chemical exposures is subject to debate. However, because it is generally unethical to experiment on humans by exposing them to known doses of chemical agents, animal toxicological evidence often provides the best scientific information about the risk of disease from a chemical exposure.⁴⁴

43. See section titled “Extrapolation from Animal (In Vivo) and Cell (In Vitro) Research to Humans” below.

44. Clinical drug trials for therapeutic drugs are an example of experimental human exposure, where as a policy matter potential benefit is considered greater than the risk of adverse events. Notably, information from animal research is used to determine risk profiles, prior to human trials. The human trials are highly regulated, and subject to ethical requirements, including informed

In contrast to their exposure to drugs, only rarely are humans exposed to environmental chemicals in a manner that permits a quantitative determination of adverse outcomes.⁴⁵ This area of toxicological study may consist of individual or multiple case reports, or even experimental studies in which individuals or groups of individuals have been exposed to a chemical under circumstances that permit analysis of dose–response relationships, mechanisms of action, or other aspects of toxicology. For example, individuals occupationally or environmentally exposed to PCBs have been studied to determine the routes of absorption, distribution, metabolism, and excretion for this group of industrial chemicals. Human exposure occurs most frequently in occupational settings where workers are often exposed each working day to industrial chemicals. In well-regulated workplaces, there is often information about the exposure levels of specific chemicals of concern to the average worker. However, even under these circumstances, determining the amount (concentration, frequency, and duration) of exposure of an individual worker is more problematic, particularly as it must be done retrospectively in most toxic tort cases. Moreover, human populations are exposed to many other chemicals and risk factors, including those associated with normal aging, making it more difficult to isolate the role of any one chemical in the increased risk of a disease.⁴⁶

Toxicologists use a wide range of experimental techniques, depending in part on their area of specialization. Toxicological research may focus on classes of chemical compounds, such as solvents, metals, and pesticides; body system effects, such as effects on the nervous system (neurotoxicology), the liver (hepatotoxicology), the reproductive system (reproductive toxicology), and the immune system (immunotoxicology); chemical effects on complex disease processes such as cancer (chemical carcinogenesis) and birth defects (developmental toxicology); and effects on physiological processes, including inhalation toxicology, cardiovascular toxicology, and dermal (skin) toxicology. Each of these subdisciplines utilizes both *in vivo* and *in vitro* research.⁴⁷ Of importance to all of these types of toxicological studies is the rapidly growing field of molecular toxicology, which examines how certain chemicals interact with biomolecules within the cell, such as proteins, lipids, and DNA. Molecular toxicology studies provide critical insights into cross-species extrapolation, dose–response, and mode of action that demonstrates the biological plausibility used for causal inference following human exposure.

consent. See, e.g., FDA, Clinical Trials Guidance Documents, <https://www.fda.gov/science-research/clinical-trials-and-human-subject-protection/clinical-trials-guidance-documents>.

45. However, it is from drug studies in which multiple animal species are compared directly with humans that many of the principles of toxicology have been developed.

46. See, e.g., OECD, Considerations for Assessing the Risks of Combined Exposure to Multiple Chemicals (2018), <https://perma.cc/L48K-W79J>.

47. See, e.g., Casarett & Doull's, *supra* note 1.

The insights provided by molecular toxicology are of particular interest to regulatory toxicologists who are exploring whether these insights can form the basis for new, more sensitive screening techniques to determine the risk of new or existing chemicals.⁴⁸

In Vivo Research: Use of Live Animals in Toxicity Testing and Safety Assessment

Animal research in toxicology generally falls under two headings: safety assessment and classic laboratory research, with a continuum between them. As explained in the section titled “Safety and Risk Assessment” above, safety assessment is a relatively formal approach in which a chemical’s potential for toxicity is tested in vivo or in vitro using standardized techniques often prescribed by regulatory agencies, such as the EPA and the FDA.⁴⁹

The roots of toxicology in the science of pharmacology are reflected in an emphasis on understanding the absorption, distribution, metabolism, and excretion of chemicals. Basic toxicological laboratory research also focuses on the mechanisms or modes of action of external chemical and physical agents.⁵⁰ Such research is based on the standard elements of scientific studies, including appropriate experimental design using control groups and statistical evaluation. In general, toxicological research attempts to hold all variables constant except for that of the chemical exposure. Any change in the experimental group not found in the control group is assumed to be caused by the chemical.

Determining Dose–Response Relationships

An important component of toxicological research is dose–response relationships. Thus, most toxicological studies generally test a range of doses of the chemical. Animal experiments are conducted to determine the dose–response relationships of a compound by measuring how response varies with dose,

48. Robert J. Kavlock et al., *Accelerating the Pace of Chemical Risk Assessment*, 31 Chem. Res. Toxicology 287 (2018), <https://doi.org/10.1021/acs.chemrestox.7b00339>.

49. See, e.g., Michael Dorato et al., *The Toxicologic Assessment of Pharmaceutical and Biotechnology Products*, in Hayes’ Principles and Methods of Toxicology 325 (A. Wallace Hayes & Claire L. Kruger eds., 6th ed. 2014), <https://doi.org/10.1201/b17359>.

50. The terms *mechanism of action* and *mode of action* are conceptually similar and are used to describe the specific molecular, biochemical, and cellular changes that are responsible for the observed toxic effects following the administration of a chemical. Although not identical, the two terms have similar meaning. Generally, mode of action is somewhat broader and refers to the generalized sequence of molecular events that lead to toxicity, whereas mechanism of action refers to a specific cellular, biochemical, or molecular target of the substance (e.g., a specific receptor).

including diligently searching for a dose that has no measurable adverse effect. This information is useful in understanding the mechanisms of toxicity and in extrapolating data from animals to humans.⁵¹

Acute Toxicity Testing and the Lethal Dose 50% (LD₅₀)

To determine the dose–response relationship for a compound, a short-term lethal dose 50% (LD₅₀) may be derived experimentally. The LD₅₀ is the dose at which a compound kills 50% of laboratory animals and usually is determined from a single exposure, although the observation may extend for a period of days to weeks. The use of this easily measured endpoint for acute toxicity largely has been replaced, in part because recent advances in toxicology have provided other pertinent endpoints, and in part because of animal welfare concerns that have focused on reduction, refinement, or replacement of the use of animals in laboratory research.⁵² Although determining the LD₅₀ was frequently the primary goal of acute toxicity testing, much additional useful information is obtained, such as the identification of the target organ(s) of effect, overall characterization of the symptoms, and progression of toxicity following single doses. Acute toxicity studies are also necessary to establish the dose range necessary for subsequent repeated-dose (subacute, usually fourteen days; and subchronic, usually ninety days) toxicity studies.

No Observed Adverse Effect Level (NOAEL)

A dose–response study also permits the determination of another important characteristic of the biological action of a chemical—the no observed adverse effect level (NOAEL; Figure 1).⁵³ The NOAEL sometimes is called a threshold, because it is the level above which observed adverse or toxic effects in test

51. See sections titled “Benchmark Dose (BMD),” “Linearized Non-Threshold (LNT) Model and Determination of Cancer Risk,” and “Maximum Tolerated Dose (MTD) and Chronic Toxicity Tests” below.

52. Dale M. Cooper et al., *The Humane Use and Care of Laboratory Animals in Toxicology Research*, in Hayes’ *Principles and Methods of Toxicology* 1023, *supra* note 49.

53. For example, undiluted acid on the skin can cause a horrible burn. As the acid is diluted to lower and lower concentrations, less and less of an effect occurs until there is a concentration sufficiently low (e.g., one drop in a bathtub of water, or a sample with less than the acidity of vinegar) that no effect occurs. This “no observed effect” concentration differs from person to person. For example, a baby’s skin is more sensitive than that of an adult, and skin that is irritated or broken responds to the effects of an acid at a lower concentration. However, the key point is that there is some concentration that is completely harmless to the skin.

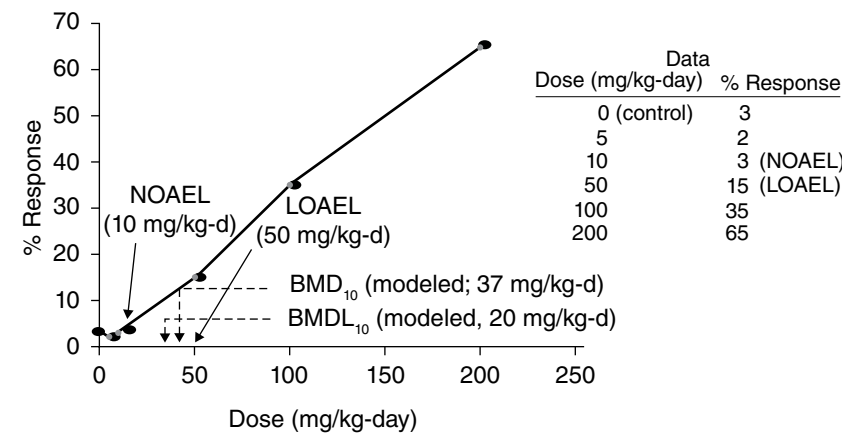
animals are believed to occur and below which no toxicity is observed.⁵⁴ Of course, because the NOAEL is dependent on the ability to observe an effect, the level is sometimes lowered after more sophisticated methods for detecting effects are developed. Typically, NOAEL studies involve the administration of the test chemical at varying different doses, daily over a period of ninety days. At least three doses (plus the “zero dose” control) are used, with doses generally spanning a dose range of thirty- to one hundred-fold, with the highest dose causing clear evidence of toxicity, and the lowest dose showing no evidence of toxicity. Typically, rats are used, with at least ten males and ten females for each dose group. The ultimate goal is to establish a dose–response curve, with at least one dose that does not show any evidence of toxicity. In circumstances where the lowest dose tested resulted in an adverse effect, the terminology for that dose is the lowest observed adverse effect level (LOAEL), which may then be used in place of a NOAEL, with additional caveats (e.g., uncertainty factors) if used in risk assessment.

Benchmark Dose (BMD)

For regulatory toxicology, the NOAEL has been replaced by a more statistically robust approach known as the benchmark dose (BMD; Figure 1). The BMD is determined based on dose–response modeling and is defined as the exposure associated with a specified low incidence of risk, generally in the range of 1% to 10%, of a health effect, or the dose associated with a specified measure or change of a biological effect (thus, if the lower range on the BMD is set at a 10% response, it is referred to as a BMD₁₀). To model the BMD, sufficient data must exist, such as at least a statistically or biologically significant dose–related trend in the

54. The significance of the NOAEL was relied on in an action alleging various health issues caused by hazardous levels of wind-borne thallium dust that escaped a military base landfill and entered plaintiff’s living environment. *Myers v. United States*, No. 02cv1349–BEN, 2014 WL 6611398 (S.D. Cal. Nov. 20, 2014), *aff’d*, 673 F. App’x 749 (9th Cir. 2016). The court ruled in favor of defendants, in part from finding persuasive evidence that the low amount of thallium found in soil at the landfill was significantly below the NOAEL. 2014 WL 6611398, at *31. *See also* *Env’t L. Found. v. Beech-Nut Nutrition Corp.*, 235 Cal. App. 4th 307, 311, 317–18 (2015), *as modified on denial of reh’g* (Apr. 16, 2015) (affirming judgment in favor of defendants because it found the trial court did not err in relying on the defendant’s toxicology expert testimony that the average and single-exposure levels of lead in the manufacturer’s baby food does not exceed the NOAEL (i.e., the regulatory safe harbor exposure level)). *But see* *Chiaracane v. Port Auth. Trans-Hudson Corp.*, No. 18-CV-2995, 2020 WL 906268, at *4, *8 (S.D.N.Y. Feb. 25, 2020) (holding that plaintiffs’ expert testimony by a neurotoxicologist on the issue of causation of respiratory symptoms by a cleaning chemical cannot be admitted because, although the expert spoke to NOAEL, the expert did not quantify the plaintiffs’ dose or exposure to the chemical in his analysis).

Figure 1. Classical dose–response relationship showing various terms associated with points of departure for use in risk assessment.



Data reflect hypothetical experiment with twenty animals at each of five dose groups and a control group:

- NOAEL, No Observed Adverse Effect Level (dose; mg/kg-d)
- LOAEL, Lowest Observed Adverse Effect Level (dose; mg/kg-d)
- BMD₁₀, Benchmark Dose at 10% response level (mg/kg-d)
- BMDL₁₀, 95th percentile Lower Confidence Limit on the BMD10 (mg/kg-d)

selected endpoint.⁵⁵ Although the NOAEL and BMD are conceptually similar, the BMD is generally considered to be a more scientifically robust value because it uses all of the dose–response information from a multidose study, rather than just the single dose point that was associated with no observable response. In the regulatory arena, the upper and lower 95% confidence limits on the BMD are often statistically determined. The lower confidence limit, called the BMDL, is often used as the point of departure (POD) for extrapolating to lower doses or determining reference doses (RfD) or acceptable daily intakes (ADIs). If the BMD₁₀ is used, then the lower bounds response is called the BMDL₁₀ (Figure 1). RfDs and ADIs are then determined by dividing the point of departure by uncertainty or safety

55. Courts also recognize the benchmark dose. *See, e.g., Wyman v. U.S. Surgical Corp.*, No. 1:18-cv-00095-JAW, 2020 WL 1932338, at *21 (D. Me. Apr. 21, 2020) (explaining that the EPA bases its estimations of daily exposure levels of methylmercury a human can have without “appreciable risk of deleterious effects during a lifetime” on benchmark doses, which are amounts of methylmercury that are “associated with the threshold for a health effect on sensitive populations according to epidemiological studies”); *Kolakowski v. Sec’y of Health & Hum. Servs.*, No. 99-0625V, 2010 WL 5672753, at *11 (Fed. Cl. Nov. 23, 2010) (explaining that the National Academy of Sciences establishes its risk assessment benchmark dose for mercury “from a percentage of the population with a physical response to the dose”).

factors. Many chemicals have multiple different toxic effects, and the dose–response relationship for each adverse outcome (toxic effect) may be different. Typically, regulatory agencies use the most sensitive toxic effect (the effect associated with the lowest dose or threshold) to establish regulatory values, such as Maximum Contaminant Levels (MCLs) under the Clean Water Act, or National Ambient Air Quality Standards (NAAQS) under the Clean Air Act. Thus, in the absence of robust epidemiological data, the NOAEL or BMDL for the most sensitive toxic effect are frequently used to establish “safe” or “acceptable” levels of exposure, which are obtained by simply dividing the POD (identified as the BMDL₁₀ or NOAEL) by safety or uncertainty factors (SF or UF). In the early days of regulation of food additives and contaminants, NOAELs were established as discussed above, and typically divided by a factor of one hundred to account for: (1) potential species differences, with a default assumption that humans might be tenfold more sensitive than the rat or mouse used to derive the NOAEL, and (2) interindividual variability within the human population, with a default assumption of another factor of ten. Today, regulatory agencies such as the EPA and the FDA often still use these basic assumptions or uncertainty factors but may modify overall uncertainty by adding additional uncertainty factors based on the quality of the study and the target population of concern. For example, the Food Quality Protection Act of 1996⁵⁶ generally requires an additional uncertainty factor (again, usually a factor of ten) if potentially susceptible populations, such as children and pregnant women, are likely to be exposed.

In toxic tort litigation related to a specific effect (i.e., a particular type of birth defect, liver or kidney disease, neurological damage, etc.), experts may argue that the dose of the agent was below the threshold and thus insufficient to have caused or substantially contributed to the disease. An individual plaintiff’s levels of exposure to a toxic substance can sometimes be estimated using exposure science techniques (see *Reference Guide on Exposure Science and Exposure Assessment*) and are usually expressed in units of milligrams (mg) of chemical per kilogram (kg) of body weight, per day. These are the same units used in determining NOAEL and BMDL levels from experimental animal studies, and sometimes from human epidemiology studies where airborne, dietary, or drinking water concentrations have been measured. Comparison of the estimated NOAEL or BMDL to the estimated dose of an individual can provide a useful approximation of whether the estimated exposure is likely to have exceeded a level known to be toxic. The ratio of the human-exposed dose to that of an estimate of a potential “threshold” (e.g., BMDL or NOAEL with appropriate safety factor adjustments) is called the Margin of Exposure (MOE). The larger the MOE, the less likely it is that an individual’s actual exposure could have caused or contributed to the disease. Thus, an MOE greater than 1,000 is unlikely to be sufficient to have caused or

56. See EPA, Summary of the Food Quality Protection Act, <https://perma.cc/5G7J-KYLS>.

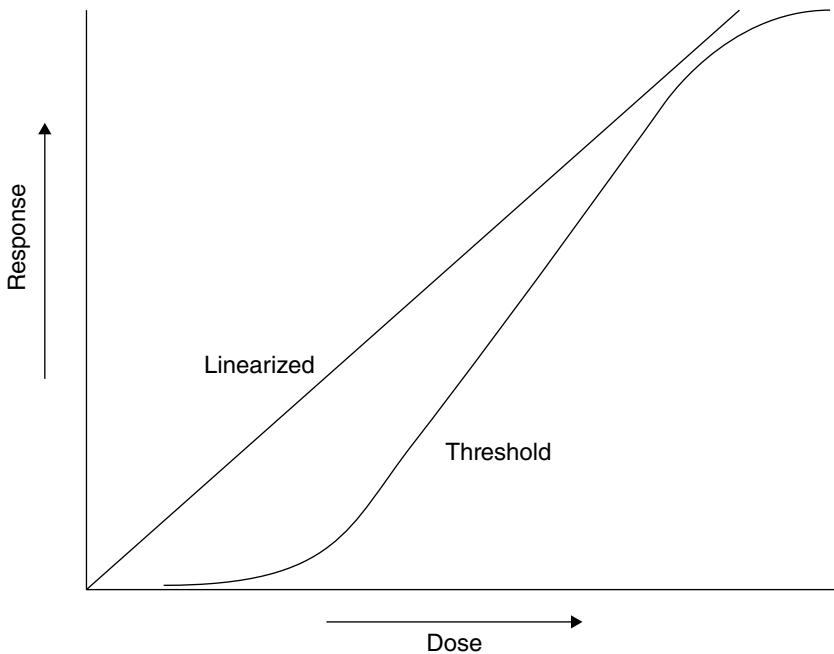
substantially contributed to a toxic effect, whereas an MOE that is less than 1.0 might reasonably be anticipated to have caused or contributed to the toxic effect in question. However, there are many assumptions that go into such estimates and comparisons, and each assumption should be justified with data. It is important to note that this approach to estimation of a potential adverse outcome (risk) is generally not used for estimating the probability of cancer risk from chemicals that are thought to cause cancer by damaging DNA—so-called genotoxic carcinogens. The reason for this is that regulatory policies for genotoxic carcinogens have generally assumed that risk is proportionately related to dose and thus there is some risk at any dose, as discussed below.

Linearized Non-Threshold (LNT) Model and Determination of Cancer Risk

Certain genetic mutations, such as those leading to cancer and some inherited disorders, have been widely hypothesized to occur without any threshold. In theory, the cancer-causing mutation to the genetic material of the cell can be produced by any one molecule of certain chemicals. The LNT model led to the development of the one-hit hypothesis of cancer risk, in which each molecule of a cancer-causing chemical has some finite possibility of producing the mutation that leads to cancer. (See Figure 2 for an idealized comparison of an LNT and threshold dose–response.) This risk is very small, because it is unlikely that any one molecule of a potentially cancer-causing agent will reach that one particular spot in a specific cell’s DNA and result in the specific change that then eludes the body’s defenses and leads to a clinical case of cancer. However, the LNT hypothesis holds that the risk is not zero. The same model also can be used to predict the risk of inheritable mutational events. There is also considerable scientific controversy over the origins of the LNT model and whether it is scientifically appropriate to use for risk assessment of chemical carcinogens.⁵⁷

57. For further discussion of the LNT model of carcinogenesis, see James E. Klaunig & Zemin Wang, *Chemical Carcinogenesis*, in Casarett & Doull’s, *supra* note 1, at 329. See also Edward J. Calabrese et al., *Dose Response and Risk Assessment: Evolutionary Foundations*, 309 *Env’t Pollution* 119787 (2022), <https://doi.org/10.1016/j.envpol.2022.119787>; Rebecca A. Clewell et al., *Dose-Dependence of Chemical Carcinogenicity: Biological Mechanisms for Thresholds and Implications for Risk Assessment*, 301 *Chemico-Biological Interactions* 112 (2019), <https://doi.org/10.1016/j.cbi.2019.01.025>. Although the one-hit model may explain the response to some carcinogens, there is accumulating evidence that for most cancers there is in fact a multistage process and that some cancer-causing agents, so-called epigenetic or nongenotoxic agents, act through non-mutational processes. For example, the multistage cancer process may explain the carcinogenicity of benzo[a]pyrene (produced by the combustion of hydrocarbons such as oil) and aflatoxin (a mold toxin that contaminates corn and peanuts). However, non-mutational responses to asbestos, dioxin, and estradiol cause their carcinogenic effects. The appropriate mathematical model to use to depict the dose–response relationship for such carcinogens is still a matter of debate. See John E. Doe et al., *The Codification of Hazard and Its Impact on the Hazard Versus Risk Controversy*, 95 *Archives Toxicology* 3611 (2021), <https://doi.org/10.1007/s00204-021-03145-6>.

Figure 2. Idealized comparison of a linearized non-threshold and threshold dose–response relationship.



Proposals have been made to merge cancer and noncancer risk assessment models. Nat'l Rsch. Council & Comm. on Improving Risk Analysis Approaches Used by the U.S. EPA, *supra* note 20. The controversy over the use of non-threshold carcinogenesis models in regulatory risk assessment is described by J.V. Rodricks, *When Risk Assessment Came to Washington: A Look Back*, 17 Dose Response 1 (2019), <https://doi.org/10.1177/1559325818824934>. Recent efforts to improve risk assessment for non-genotoxic carcinogens, which are generally not thought to follow low dose, linear response, are underway. See Christian Strupp et al., *Increased Cell Proliferation as a Key Event in Chemical Carcinogenesis: Application in an Integrated Approach for the Testing and Assessment of Non-Genotoxic Carcinogenesis*, 24 Int'l J. Molecular Scis. 13246 (2023), <https://doi.org/10.3390/ijms241713246>.

Courts continue to grapple with the non-threshold model. The District of Columbia Court of Appeals noted “‘mounting judicial skepticism’ of no-threshold models. . . .” *N. Am.’s Bldg. Trades Unions v. Occupational Safety & Health Admin.*, 878 F.3d 271, 284 (D.C. Cir. 2017). Nonetheless, the court upheld OSHA’s use of a non-threshold model to determine risk for silicosis and lung cancer from silica exposure as “supported by substantial evidence” and “in line with our precedent.” *Id.* at 283–84. In another case, the District of Massachusetts excluded an expert’s opinion based on a non-threshold model, noting that such an “opinion is inadmissibly unreliable” because “there is no scientific evidence that the linear no-safe threshold analysis is an acceptable scientific technique” to determine specific causation. *Milward v. Acuity Specialty Prods. Grp., Inc.*, 969 F. Supp. 2d 101, 110 (D. Mass. 2013), *aff’d sub nom.* *Milward v. Rust-Oleum Corp.*, 820 F.3d 469, 474–75 (1st Cir. 2016). *But see* *McManaway v. KBR, Inc.*, No. H-10-1044, 2012 WL 13059744, at *11–12 (S.D. Tex. Aug. 22, 2012) (permitting expert to use non-threshold model despite an argument that the “theory . . . is unsupported by relevant case law and in the relevant scientific community” because the Fifth Circuit has “no bright line” rule excluding the model).

Maximum Tolerated Dose (MTD) and Chronic Toxicity Tests

Another type of study uses different doses of a chemical agent to establish over a ninety-day period what is known as the maximum tolerated dose (MTD), the highest dose that does not cause significant overt toxicity. The MTD is important because it enables researchers to calculate the dose of a chemical to which an animal can be exposed without reducing its lifespan, thus permitting the evaluation of the chronic effects of exposure.⁵⁸ These studies are designed to last the lifetime of the species.

Chronic toxicity tests evaluate carcinogenicity or other types of toxic effects. Federal regulatory agencies frequently require carcinogenicity studies on both sexes of two species, usually rats and mice, administered daily over the lifetime (generally two years) of the animals. A pathological evaluation is done on the tissues of animals that died during the study and those that are sacrificed at the conclusion of the study.

The rationale for using the MTD in chronic toxicity tests, such as carcinogenicity bioassays, often is misunderstood. It is preferable to use realistic doses of carcinogens in all animal studies. However, this leads to a loss of statistical power, thereby limiting the ability of the test to detect carcinogens or other toxic compounds. Consider the situation in which a realistic dose of a chemical causes a tumor in one in one hundred laboratory animals. If the lifetime background incidence of tumors in animals without exposure to the chemical is six in one hundred, a toxicological test involving one hundred control animals and one hundred exposed animals who were fed the realistic dose would be expected to reveal six control animals and seven exposed animals with the cancer. This difference is too small to be recognized as statistically significant. However, if the study started with ten times the realistic dose, the researcher would expect to get ten additional cases, for a total of sixteen cases in the exposed group and six cases in the control group, a statistically significant difference that is unlikely to be overlooked.

Unfortunately, even this example does not demonstrate the difficulties of determining risk. Regulators are responding to public concern about cancer by regulating risks often as low as one in 1,000,000—not one in one hundred, as in the example given above. To test risks of one in 1,000,000, a researcher would have to either increase the lifetime dose from ten times to 100,000 times the realistic dose or expand the numbers of animals under study into the millions. However, increases of this magnitude are beyond the world's animal testing capabilities and are also prohibitively expensive. Inevitably, then, animal studies

58. Even the determination of the MTD can be fraught with controversy. See Environmental Science Deskbook § 4:5 (2022) (illustrating problems with making inferences about human toxicity based on MTD established by animal studies).

must trade statistical power for extrapolation from higher doses to lower doses. However, recent research indicates that the MTD used in studies of some chemicals—and perhaps even the lower doses—may overwhelm biologically important protective pathways, thereby causing cellular damage integral to the disease process that would not occur at lower, more realistic doses to which humans may be exposed. Because of these concerns, there has been increasing research interest among toxicologists to redefine the MTD to include assessment of repair response pathways, where data are available, such as DNA repair processes, to ensure that the highest dose used does not overwhelm protective pathways that are operable at lower doses.⁵⁹

Accordingly, proffered toxicological expert opinion on potentially disease-causing chemicals almost always is based on evaluation of the research studies that extrapolate from animal experiments involving doses significantly higher than that to which humans are exposed.⁶⁰ In both regulatory and toxic tort scenarios, such extrapolation is backed up by other types of data, where available, such as epidemiological studies.

In Vitro Research

In vitro research concerns the effects of a chemical on human or animal cells, bacteria, yeast, isolated tissues, or embryos. Thousands of in vitro toxicological tests have been described in the scientific literature. Many tests are for chemicals that cause changes in DNA (mutations) in bacterial or mammalian systems. There are short-term in vitro tests for just about every physiological response and every organ system, such as perfusion tests and DNA studies. Many of these short-term tests have been thoroughly validated, and standardized protocols for conducting such studies are widely available.⁶¹ However, in vitro studies often do not

59. Yu-Mei Tan et al., *Opportunities and Challenges Related to Saturation of Toxicokinetic Processes: Implications for Risk Assessment*, 127 Regul. Toxicology & Pharmacology 105070 (2021), <https://doi.org/10.1016/j.yrtph.2021.105070>.

60. See, e.g., Int'l Agency for Research on Cancer, World Health Organization, Preamble, Amended 2019, <https://perma.cc/A3L5-NLN7>; Joseph V. Rodricks, *Evaluating Disease Causation in Humans Exposed to Toxic Substances*, 14 J.L. & Pol'y 39 (2006).

61. See Joanna Klapacz & B. Bhaskar Gollapudi, *Genetic Toxicology*, in Casarett & Doull's, *supra* note 1, at 497–54. Use of in vitro data for evaluating human mutagenicity and teratogenicity is described in John M. Rogers, *Developmental Toxicology*, in Casarett & Doull's, *supra* note 1, at 547–92. For a critique of expert testimony using in vitro data, see *In re Zolofit (Sertraline Hydrochloride) Prods. Liab. Litig.*, 26 F. Supp. 3d 466, 477–82 (E.D. Pa. 2014); *In re Mirena IUS Levonorgestrel-Related Prods. Liab. Litig. (No. II)*, 341 F. Supp. 3d 213, 294–95 (S.D.N.Y. 2018), *aff'd*, 982 F.3d 113 (2d Cir. 2020); *Davis v. McKesson Corp.*, No. CV-18-1157-PHX-DGC, 2019 WL 3532179, at *17 (D. Ariz. Aug. 2, 2019); *In re Incretin-Based Therapies Prods. Liab. Litig.*, 524 F. Supp. 3d 1007 (S.D. Cal. 2021), *aff'd*, No. 21-55342, 2022 WL 898595 (9th Cir. Mar. 28, 2022).

accurately reflect the complexity of in vivo biological responses to chemicals, and thus can be misleading. A particular challenge with interpreting in vitro studies is how to compare the concentration of the test chemical used in the test tube or petri dish with the actual concentration of the same chemical in target tissues of the individual exposed to the chemical (the in vivo situation). As such, in vitro studies are often “hypothesis generating,” rather than providing definitive scientific proof that a chemical can cause a particular type of toxic response in vivo.

The criteria of reliability for an in vitro test include the following: (1) whether the test has come through a published protocol in which many laboratories used the same in vitro method on a series of unknown compounds prepared by a reputable organization (such as the NIH or the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use [ICH]) to determine if the test consistently and accurately measures toxicity; (2) whether the test has been adopted by a U.S. or international regulatory body; and (3) whether the test is predictive of in vivo outcomes related to the same cell or target organ system.

Understanding the molecular pathways that are responsible for the ability of a chemical to cause a particular toxic effect, such as a specific type of cancer, is of value to a toxicologist charged with assessing the likelihood that a specific chemical caused a specific cancer in an individual. Identification of biological pathways is an important component of specific causation in most toxic tort cases. Thus, in vitro toxicology studies that investigate exactly how it is that a chemical causes its adverse biological effects are relevant to toxic tort litigation and can contribute significant support to a toxicologist’s expert opinions on specific causation.⁶²

New Approach Methodologies (NAMs)

Advances in analyzing the impact of external chemical agents are dependent primarily on advances in biology and in medicine leading to a greater understanding of normal body functions and of the changes observed in disease states. Integrating advances in biological and medical understanding of the cellular, subcellular, and molecular basis for normal function is reflected in the continued incorporation of new multidisciplinary science into toxicology.

62. See, e.g., *In re Johnson & Johnson Talcum Powder Prods. Mktg., Sales Pracs. & Prods. Liab. Litig.*, 509 F. Supp. 3d 116 (D.N.J. 2020) (expert may opine that his in vitro studies show that talc causes cellular inflammation and oxidative stress but may not offer an opinion on a causal link between talc and ovarian cancer); see also *In re Denture Cream Prods. Liab. Litig.*, No. 09-2051-MD, 2015 WL 392021 (S.D. Fla. Jan. 28, 2015) (expert’s failure to explain the relevancy of in vitro studies to humans or to account for factors needed to make a proper extrapolation renders expert’s opinion unreliable); *Hardeman v. Monsanto Co.*, 997 F.3d 941, 963 (9th Cir. 2021) (“[C]ell studies can support more substantial evidence of causation.”).

The driving forces for incorporating these newer advances into the field of toxicology have included the recognition that a large percentage of chemicals in commerce had been inadequately tested for toxicity. The resultant changes in laws in the European Union (REACH), the United States (the Lautenberg Act), and elsewhere have led to development and adoption of new approach methodologies (NAMs) that could replace standard animal toxicology.⁶³ Pressure toward the use of subcellular and computer-based *in silico* methodologies has also come from animal rights proponents. Pharmaceutical companies also have invested in developing *in silico* technology for rapid survey of the potential biological impact of large numbers of chemicals to support their search for therapeutic agents without unanticipated side effects. NAMs have also considered how to more effectively incorporate individual and population exposure data into chemical evaluation.

The NAMs collectively have a distance to go before they can be fully accepted as pertinent to regulatory decisions or to toxic torts. Important issues include understanding whether a gene or a biological pathway that is turned on by a chemical in a test tube or *in silico* will result in a similar response *in vivo*; uncertainties about the implications of the response to human or ecosystem function; the importance of reproducing the wide range of human sensitivity and environmental variability; and the potential difficulty of distinguishing between a pathway to an adverse effect in contrast to a desirable response that counteracts the potential for an adverse effect. Dose–response analysis is also problematic, as it is usually highly uncertain how a concentration of a chemical in a test tube or cell culture flask relates to a blood concentration following an *in vivo* exposure. Comparison of the relative efficacy of these newer methodologies will determine whether they will be incorporated into expert opinions in tort or regulatory law cases.

Extrapolation from Animal (In Vivo) and Cell (In Vitro) Research to Humans

Two types of extrapolation must be considered: from animal data to humans and from higher doses to lower doses.⁶⁴ In qualitative extrapolation, one can usually rely on the fact that a compound causing an effect in one mammalian species will cause it in another species. This is a basic principle of toxicology and pharmacology, when properly qualified by known differences in absorption, metabolism, excretion, and other known biological differences (e.g., differences in

63. Nat'l Rsch. Council, *Toxicity Testing in the 21st Century: A Vision and a Strategy* (2007), <https://doi.org/10.17226/11970>. See also Kavlock et al., *supra* note 48; Sebastian Schmeisser et al., *New Approach Methodologies in Human Regulatory Toxicology—Not If, but How and When!*, 178 *Env't Int'l* 108082 (2023), <https://doi.org/10.1016/j.envint.2023.108082>.

64. See EPA, *Guidance for Applying Quantitative Data to Develop Data-Derived Extrapolation Factors for Interspecies and Intraspecies Extrapolation* (2014), <https://perma.cc/JJ2K-SD8C>.

receptor binding affinity) between humans and experimental animals. If a heavy metal, such as mercury, causes kidney toxicity in laboratory animals, it is highly likely to do so at some dose in humans. However, the dose at which mercury causes this effect in laboratory animals is modified by many internal factors, and the dose–response curve may be different from that for humans. Through the study of factors that modify the toxic effects of chemicals, including absorption, distribution, metabolism, and excretion, researchers can improve the ability to extrapolate from laboratory animals to humans and from higher to lower doses.⁶⁵ The mathematical depiction of the process by which an external exposure is absorbed into the bloodstream and then moves through various compartments in the body until it reaches the target organ is often called physiologically based pharmacokinetic or toxicokinetic (PBPK/PBPT) modeling.⁶⁶

Extrapolation from studies in nonmammalian species to humans is much more difficult but can be done if there is sufficient information on similarities in absorption, distribution, metabolism, and excretion. Advances in computational toxicology have increased the ability of toxicologists to make such extrapolations.⁶⁷ Quantitative determinations of human toxicity based on *in vitro* studies usually are not considered appropriate. *In vitro* or *in vivo* animal data for elucidating the mechanisms of toxicity are more persuasive when positive human epidemiological data or human tissue–based toxicological information also exists.

65. For example, benzene undergoes a complex metabolic sequence that results in toxicity to the bone marrow in all species, including humans. Martyn T. Smith, *Advances in Understanding Benzene Health Effects and Susceptibility*, 31 Annual Rev. Public Health 133 (2010), <https://doi.org/10.1146/annurev.publhealth.012809.103646>. The exact metabolites responsible for this bone marrow toxicity are the subject of much interest but remain unknown. Mice are more susceptible to benzene than are rats. If researchers could determine the differences between mice and rats in their metabolism of benzene, they would have a useful clue about which portion of the metabolic scheme is responsible for benzene toxicity to the bone marrow. See, e.g., Angela L. Slitt, *Absorption, Distribution, and Excretion of Toxicants*, in Casarett & Doull’s, *supra* note 1, at 159–92; Andrew Parkinson et al., *Biotransformation of Xenobiotics*, in Casarett & Doull’s, *supra* note 1, at 193–400.

66. For a discussion of methods used to extrapolate from animal toxicity data to human health effects, see text and references cited in the sections titled “Benchmark Dose (BMD)” and “Linearized Non-Threshold (LNT) Model and Determination of Cancer Risk” above.

67. See, e.g., Nicole C. Kleinstreuer et al., *Introduction to Special Issue: Computational Toxicology*, 34 Chem. Res. Toxicology 171 (2021), <https://doi.org/10.1021/acs.chemrestox.1c00032>. For discussion of the use of artificial intelligence and machine learning in toxicology, see Zhoumeng Lin & Wei-Chun Chou, *Machine Learning and Artificial Intelligence in Toxicological Sciences*, 189 Toxicological Scis. 7 (2022), <https://doi.org/10.1093/toxsci/kfac075>. For an overview of new approaches in structure–activity relationships, see Mark T.D. Cronin et al., *A Scheme to Evaluate Structural Alerts to Predict Toxicity—Assessing Confidence by Characterising Uncertainties*, 135 Regul. Toxicology & Pharmacology 105249 (2022), <https://doi.org/10.1016/j.yrtph.2022.105249>.

Toxicological Processes and Target Organ Toxicity

The biological, chemical, and physical phenomena that are the basis of life are astounding in their complexity. As a result, human subcellular, cellular, and organ function are both delicately balanced and highly robust. Small changes caused by external chemical and physical agents can have major effects; yet, through the millennia, evolutionary pressures have led to the emergence of safety mechanisms that defend against adverse environmental stresses.

The specialization that is a hallmark of organ development in vertebrates inherently leads to diversity in the underlying processes that are the basis of organ function. Certain chemicals poison virtually all cells by affecting a basic biological process essential to life. For example, cyanide interferes with the conversion of oxygen to energy in a subcellular component known as mitochondria.⁶⁸ Other chemical agents interfere selectively with an organ-specific process. For example, organophosphorus pesticides and chemically related biological warfare agents (i.e., some nerve gases) specifically interfere with the way nerve cells transmit signals—a process that is pertinent primarily to the nervous system. Other chemicals, such as the widely used over-the-counter analgesic acetaminophen, can cause specific damage to the liver if the dose is sufficient to overwhelm the normal protective pathways in the liver.⁶⁹ Table 1 provides arbitrarily selected examples of toxicological endpoints and agents of concern, which are not meant to be inclusive or exhaustive.

Despite this specialization, there are pathological processes common to diseases affecting many different organs. For example, chronic inflammation of the skin leads to fiber formation that is recognized as scarring. Similarly, cirrhosis of the liver can result from fibrogenic processes caused by repetitive inflammation of the liver, such as from the overuse of ethanol, and fibrosis of the lung is a pathological process resulting from exposure to asbestos, silica, and

68. Note that the diffuse toxicity of cyanide also reflects its ability to spread widely in the body. Certain mitochondrial poisons primarily affect the brain and active muscles, including the heart, which are particularly oxygen dependent. Others, unable to penetrate the blood–brain barrier, will primarily affect peripheral muscles including the heart.

69. Acetaminophen (TylenolTM) is among the most widely used analgesic/non-steroidal anti-inflammatory drugs, and is used safely every day by millions of people at normal therapeutic doses (less than five grams per day). However, doses in excess of ten to fifteen grams in a single day deplete an intracellular antioxidant called glutathione (GSH), and the excess dose can cause serious damage to the liver. Acetaminophen overdose is among the leading causes of acute liver failure in the United States. To produce toxicity it must first be metabolized to a reactive form by a specific enzyme that is only present in significant amounts in the liver. This is why acetaminophen is only toxic to the liver.

Table 1. Sample of Selected Toxicological Endpoints and Examples of Agents of Concern in Humans

Organ System	Examples of Endpoints	Examples of Agents of Concern
Skin	allergic contact dermatitis	nickel, poison ivy, cutting oils
	chloracne	dioxins and dioxin-like compounds
	cancer	polycyclic aromatic hydrocarbons
Respiratory Tract	nonspecific irritation (reactive airway disease)	formaldehyde, acrolein, ozone
	asthma	toluene diisocyanate
	chronic obstructive pulmonary disease	cigarette smoke
	fibrosis, pneumoconiosis	silica, mineral dusts, cotton dust, beryllium
	cancer	cigarette smoke, arsenic, asbestos, nickel
Blood and the Immune System	anemia	arsine, lead, methyldopa
	secondary polycythemia	cobalt
	methemoglobinemia	nitrites, aniline dyes, diaminodiophenyl sulfone
	pancytopenia	benzene, radiation, chemotherapeutic agents
	secondary lupus erythematosus	hydralazine
	leukemia	benzene, radiation, chemotherapeutic agents
Liver and Gastro-intestinal Tract	hepatic damage (hepatitis)	acetaminophen, ethanol, carbon tetrachloride, vitamin A
	cancer	aflatoxin, vinyl chloride
Urinary Tract	kidney toxicity	ethylene and diethylene glycols, lead, melamine, aminoglycoside antibiotics
	bladder cancer	aromatic amines, arsenic
Nervous System	nervous system toxicity	cholinesterase inhibitors, mercury, lead, n-hexane, bacterial toxins (botulinum, tetanus), many organochlorine insecticides
	Parkinson's disease/ Parkinson's-like symptoms	manganese, MPTP

Table 1. Continued

Organ System	Examples of Endpoints	Examples of Agents of Concern
Reproductive and Developmental Toxicity	fetal malformations decreased male fertility	thalidomide, ethanol dibromochloropropane (DBCP)
Endocrine System	thyroid toxicity	radioactive iodine, perchlorate
Cardiovascular System	heart toxicity high blood pressure arrhythmias	anthracyclines, cobalt lead plant glycosides (e.g., digitalis)

Note: This table presents only examples of toxicological endpoints and examples of agents of concern in humans and is provided to help illustrate the variety of toxic agents and endpoints. It is not an exhaustive or inclusive list of organs, endpoints, or agents. Absence from this list does not indicate that any specific chemical is not a cause of the target organ toxicity.

other agents.⁷⁰ The potential for endocrine disruption by chemicals, particularly those that persist within the body, has become an increasing concern. Many of these persistent agents belong to families of chemically similar compounds, such as dioxins, PCBs, or per- and polyfluoroalkyl substances (PFAS) that may differ in their effect. Judges may encounter cases in which a wide variety of toxic effects are alleged to have been caused by endocrine-disrupting chemicals. Possible adverse outcomes from endocrine disruption include loss of fertility, birth defects, metabolic diseases that may result from thyroid disruption, and potentially some forms of cancer. While the concept of dose–response is still applicable, there is some evidence that some toxic effects of endocrine-disrupting chemicals may appear at lower doses but not be seen at higher doses, usually because the higher doses produce biological changes that mask the effects at lower doses (so-called “non-monotonic” dose–response relationship). This is an area of considerable controversy in toxicology. Vitamins also provide an example of differing adverse effects at levels above or below the normal range: e.g., too much vitamin A can cause liver disease and too little will cause blindness.

Particularly challenging to standard toxicological approaches are agents that react with different receptors present on the surface or internal components of the cell. These receptors often belong to complex families of related cellular components that are continually interacting with the broad range of hormones produced by our bodies.⁷¹ The intricate dynamic processes of normal endocrine

70. Lung fibrosis is a key pathological finding in a group of diseases known as pneumoconiosis that includes coal miners’ black lung disease, silicosis, asbestosis, and other conditions usually caused by occupational exposures to small airborne particles.

71. As a simplification, agent–receptor interactions often are described as a key in a lock, with the key needing to be able to both fit into the lock and turn the mechanism. An example from the

activity include daily cycles controlled by internal “biological clocks,” and feedback loops that allow cyclic variation, such as in the menstrual cycle or in the variations of hormone and receptor levels that are linked to normal functions such as sleeping and sexual activity. These complex normal up-and-down variations produce conceptual difficulties when attempting to extrapolate the results from model systems to the functioning human.⁷²

The understanding of the relationship between mutation and cancer discussed previously led to some of the first toxicological tests to determine whether an external agent could cause cancer. Such tests have grown in sophistication because of the advances in molecular biology and computational toxicology that have occurred concomitantly with an increased understanding of the variety of potential pathways that lead to mutagenesis.⁷³

Toxicological testing for chemical carcinogens ranges from relatively simple studies to determine whether the substance is capable of producing bacterial mutations to observation of cancer incidence as a result of long-term administration of the substance to laboratory animals. Between these two extremes are a multiplicity of tests that build upon the understanding of the mechanism of cancer causation. In vitro or in vivo tests may focus on the evidence of effects in DNA, such as the presence of adducts of the chemical or its metabolites bound to the DNA molecule or the cross-linking of the DNA molecule to protein. Researchers may look for changes in the nucleus of the cell suggestive of DNA damage that could result in mutagenesis and carcinogenesis, for example, the micronucleus test or the comet assay. Certain mutagens cause an increase in the normal exchange of nuclear material among DNA components during normal cell division, which gives rise to a test known as the sister chromatid exchange.⁷⁴ The direct observation of chromosomes

nervous system is the use in treating a heroin overdose of another opiate that has a much higher affinity for the receptor site but produces little effect once bound. When given to a normal person, this second opiate would have a mild depressant effect, but it can reverse a near fatal overdose of heroin by displacing the heroin from the receptor site. Thus, the directionality of opiate effect depends on the interaction of the components of the mixture. This interaction is even more complex when dealing with estrogenic agents that are naturally occurring as well as made within the body at different levels in response to different external and internal stimuli and at different time intervals.

72. See, e.g., *In re Denture Cream Prods. Liab. Litig.*, No. 09-2051-MD, 2015 WL 392021 (S.D. Fla. Jan. 28, 2015) (expert’s failure to correct for circadian rhythm fluctuations in zinc plasma levels rendered his opinion inadmissible). In another example, the complexity of the interaction of a mixture of dioxins with receptors governing the endocrine system can be contrasted with that of the reaction of carbon monoxide with the hemoglobin oxygen receptor discussed *supra*, note 18. The latter is unidirectional in that any additional carbon monoxide will interfere with oxygen delivery, of which there cannot be too much under normal physiological conditions.

73. See, e.g., James E. Klaunig, *Carcinogenesis*, in *An Introduction to Interdisciplinary Toxicology: From Molecules to Man* 87 (2020), <https://doi.org/10.1016/B978-0-12-813602-7.00008-9>.

74. All of these tests require validation regarding their relevance to predicting human carcinogenesis, as well as to their technical reproducibility. See Klapacz & Gollapudi, *supra* note 61, for a discussion of the wide variety of short-term mutagenesis assays for predicting mutagenic and potentially carcinogenic potential of chemicals.

to look for specific abnormalities, known as cytogenetic analysis, is providing more information about the pathways of carcinogenesis. For cancers such as acute myelogenous leukemia, it has long been recognized that those individuals who present with recognizable chromosomal abnormalities are more likely to have been exposed to a known human chemical leukemogen such as benzene.⁷⁵ These and other tests provide information that can be used in evaluating whether a chemical is a potential human carcinogen.

The recent almost explosive growth in the ability to measure minuscule changes in the complex genome, a process known as next-generation sequencing, is rapidly being applied to a broad range of human diseases. The emphasis thus far has been on the promise of developing new therapies appropriate to personalized medicine, but the same techniques may in the future provide a fingerprint of the interaction of a chemical with DNA. However, there are very few chromosomal abnormalities or specific mutations that are unequivocally linked to a specific chemical or physical carcinogen.⁷⁶

Assessing whether a chemical or physical agent produces human cancer requires careful evaluation. The World Health Organization's IARC and the U.S. National Toxicology Program (NTP) have formal processes to evaluate the weight of evidence that a chemical causes cancer.⁷⁷ Each classifies chemicals on the basis of epidemiological evidence, toxicological findings in laboratory animals, and mechanistic considerations, and then assigns a specific category of carcinogenic potential to the individual chemical or exposure situation (e.g., employment as a painter).⁷⁸ Only a small percentage of the total chemicals in commerce

75. Patrick J. Kerzic & Richard D. Irons, *Distribution of Chromosome Breakpoints in Benzene-Exposed and Unexposed AML Patients*, 55 *Env't Toxicology & Pharmacology* 212 (2017), <https://doi.org/10.1016/j.etap.2017.08.033>.

76. See Luoping Zhang et al., *The Nature of Chromosomal Aberrations Detected in Humans Exposed to Benzene*, 32 *Critical Revs. Toxicology* 1 (2002), <https://doi.org/10.1080/20024091064165>. However, studies on liver cancer patients exposed to the potent dietary carcinogen aflatoxin B1 (AFB) have demonstrated that a specific mutation in codon 249 in the important cancer gene called TP53 is uniquely associated with exposure to AFB. Only liver tumors from patients with a history of AFB exposure showed this specific mutation. S. P. Hussain et al., *TP53 Mutations and Hepatocellular Carcinoma: Insights into the Etiology and Pathogenesis of Liver Cancer*, 26 *Oncogene* 2166 (2007), <https://doi.org/10.1038/sj.onc.1210279>.

77. The U.S. National Toxicology Program issues a congressionally mandated report on carcinogens. *Nat'l Toxicology Prog., 15th Rep. on Carcinogens* (2021), <https://perma.cc/V4DP-3YNZ>. IARC produces its reports through a monograph series that provides detailed description of the agents or processes under consideration as well as the findings of the IARC expert working group. See the IARC website for a list of these monographs, <https://perma.cc/UEP2-5L9A>.

78. IARC uses the following classifications:

- Group 1, The agent (mixture) is carcinogenic to humans;
- Group 2A, The agent (mixture) is probably carcinogenic to humans;
- Group 2B, The agent (mixture) is possibly carcinogenic to humans;
- Group 3, The agent (mixture) is not classifiable as to its carcinogenicity to humans.

are considered to be known human carcinogens, although this must be considered in the context of how many of these chemicals have been rigorously tested (see discussion of REACH and the Lautenberg Act above). In the past, when chemicals were evaluated for carcinogenicity, assignment to the highest category was dependent almost totally on whether epidemiological research had been conducted, although animal data and mechanistic information were also considered. In recent years, with improved understanding of the mechanism of action of chemical carcinogens, there has been increased use of mechanistic data.⁷⁹ For example, higher credence is given to the likelihood that a chemical is a human carcinogen if the metabolite found to be responsible for carcinogenesis in a laboratory animal is also found in the blood or urine of humans exposed to this chemical, or if there is evidence of the same type of DNA damage in humans as there is in laboratory animals in which the agent does cause cancer.⁸⁰

Inherent in putting chemicals into distinct categories when there is a continuum for the strength of the evidence is that some chemicals will be very close to the dividing line between the discrete categories. Inevitably, small differences in the interpretation of the evidence for such chemicals will lead to disagreement regarding categorization.

79. See Daniele S. Wikoff et al., *A Framework for Systematic Evaluation and Quantitative Integration of Mechanistic Data in Assessments of Potential Human Carcinogens*, 167 *Toxicological Scis.* 322 (2019), <https://doi.org/10.1093/toxsci/kfy279>, for a discussion and for specific examples of the use of mechanistic data in evaluating carcinogens. The evolution in the approach to determining cancer causality is evident from reviewing the guidelines used to assemble the weight of evidence for causality by IARC and NTP, two of the organizations that have the lengthiest track record of responsibility for the hazard identification of carcinogens. Both have increased the weight given to mechanistic evidence in characterizing the overall strength of the total evidence used to classify the potential for a chemical or an exposure to be causal. IARC now permits classification in Group 1 when there is less than sufficient evidence in humans but sufficient evidence in animals and “strong evidence in exposed humans that the agent exhibits key characteristics of carcinogens and sufficient evidence of carcinogenicity in experimental animals.” IARC Preamble, *supra* note 60, at 23. The criterion used by NTP for listing a chemical as a known human carcinogen in its biannual Report on Carcinogens is: “There is sufficient evidence of carcinogenicity from studies in humans,* which indicates a causal relationship between exposure to the agent, substance, or mixture, and human cancer.” The asterisk is particularly notable in that it specifies that the evidence need not be solely epidemiological: “*This evidence can include traditional cancer epidemiology studies, data from clinical studies, and/or data derived from the study of tissues or cells from humans exposed to the substance in question, which can be useful for evaluating whether a relevant cancer mechanism is operating in humans.” See *Nat’l Toxicology Prog.*, *supra* note 77, at 6. For a recent discussion on using mechanism/mode of action data in assessment of genotoxic carcinogens, see also Andrea Hartwig et al., *Mode of Action-Based Risk Assessment of Genotoxic Carcinogens*, 94 *Archives Toxicology* 1787 (2020), <https://doi.org/10.1007/s00204-020-02733-2>. The EPA also considers mechanism of action in its regulatory approaches and distinguishes further between mechanism of action and mode of action. See Kathryn Z. Guyton et al., *Improving Prediction of Chemical Carcinogenicity by Considering Multiple Mechanisms and Applying Toxicogenomic Approaches*, 681 *Mutation Resch.* 230, 240 (2009), <https://doi.org/10.1016/j.mrrev.2008.10.001>.

80. An example is the IARC evaluation of formaldehyde that upgraded the categorization from 2A to 1 based on epidemiological data that were strongly supported by the finding of nasal cancer in laboratory animals and by the presence of DNA-protein cross-links in the nasal tissue of the

Demonstrating an Association Between Exposure and Risk of Disease

In toxic tort cases, once an expert in toxicology has been qualified, they are expected to offer an opinion on whether the plaintiff's disease was caused by exposure to a chemical. To do so, the expert relies on the principles of toxicology to provide a scientifically valid methodology for establishing causation and then applies the methodology to the facts of the case.

An opinion on causation should be premised on three preliminary assessments. First, the expert should analyze whether the disease can be related to chemical exposure by a biologically plausible hypothesis. Second, the expert should examine whether the plaintiff was exposed to the chemical in a manner that can lead to absorption into the body. Third, the expert should offer an opinion about whether the dose to which the plaintiff was exposed is sufficient to cause the disease. In complex cases, multiple experts may offer opinions as to specific aspects of disease causation.

The following questions help evaluate the strengths and weaknesses of toxicological evidence.

On What Species of Animals Was the Compound Tested? What Is Known About the Biological Similarities and Differences Between the Test Animals and Humans? How Do These Similarities and Differences Affect the Extrapolation from Animal Data in Assessing the Risk to Humans?

All living organisms share a common biology that leads to marked similarities in the responsiveness of subcellular structures to toxic agents. Among mammals, more than sufficient common organ structure and function readily permit the extrapolation from one species to another in most instances. Comparative information concerning factors that modify the toxic effects of chemicals, including

laboratory animals and of humans inhaling formaldehyde. However, epidemiological evidence associating formaldehyde with human acute myelogenous leukemia was questioned on the basis of the lack of mechanistic evidence, including questions about how such a highly reactive agent could reach the bone marrow following inhalation. See *Formaldehyde, 2-Butoxyethanol and 1-tert-Butoxypropan-2-ol*, in 88 IARC Monographs on the Evaluation of Carcinogenic Risks to Humans (2006).

absorption, distribution, metabolism, and excretion, enhances the expert's ability to extrapolate from laboratory animals to humans.⁸¹

The expert should review similarities and differences between the animal species in which the compound has been tested and humans. This analysis should form the basis of the expert's opinion regarding whether extrapolation from animals to humans is warranted.

In general, an overwhelming similarity is apparent in the biology of all living things, and there is a particularly strong similarity among mammals. Of course, laboratory animals differ from humans in many ways. For example, rats do not have gallbladders. Thus, rat data would not be pertinent to the possibility that a compound produces human gallbladder toxicity.⁸² The life stage at which exposure occurs can also be an important determinant of toxicity. It is generally recognized that the developing embryo and fetus are unusually susceptible to many toxic substances, and thus exposure of a pregnant animal requires special consideration for toxic effects that might occur to the developing offspring. In a similar manner, it is also generally recognized that early life exposure (from infancy through adolescence) is also a period of enhanced susceptibility, and effects not seen in adults may occur. This is especially true for chemicals that affect the nervous system, since the development of the nervous system is not complete until early adulthood. Thus, for example, the effects of lead poisoning in children are manifest mostly through impacts on the central nervous system that lead to impairment of learning and development, whereas adults exposed to the same level of lead have little or no detectable effects on the central nervous system, although at high doses lead can cause effects on the peripheral nervous system in adults (so-called "lead palsy").

Note that, in animal studies, many subjective symptoms are poorly modeled regardless of the age of exposure. Thus, complaints that a chemical has caused nonspecific symptoms, such as nausea, headache, and weakness, for which there may be no objective manifestations in humans, are difficult to test in laboratory animals.

81. See generally *supra* note 3 and references therein.

82. See, e.g., *Hardeman v. Monsanto Co.*, 997 F.3d 941, 963 (9th Cir. 2021) ("Animal studies are relevant evidence of causation where there is a sound basis for extrapolating conclusions from those studies to humans in real-world conditions."). See generally David Spurgeon et al., *Species Sensitivity to Toxic Substances: Evolution, Ecology and Applications*, 8 *Frontiers Env't Sci.* 1 (2020), <https://doi.org/10.3389/fenvs.2020.588380>. Species differences that produce a qualitative difference in response to xenobiotics are well known. Sometimes understanding the mechanism underlying the species difference can allow one to predict whether the effect will occur in humans. Thus, carbaryl, an insecticide commonly used for gypsy moth control, produces fetal abnormalities in dogs but not in hamsters, mice, rats, and monkeys. Dogs lack the specific enzyme involved in metabolizing carbaryl; the other species tested all have this enzyme, as do humans. Therefore, it has been assumed that humans are not at risk for fetal malformations produced by carbaryl.

Another important variable in extrapolating laboratory animal data to humans is that animals used in experimental research are maintained on nutritionally adequate diets and clean, healthy environments that are designed to maintain optimal health. The diets are free of other chemical contaminants, and the rest of their environment is equally pristine. This is in contrast to humans, who may have unhealthy diets and may be exposed to multiple different toxic substances from their diets, workplace, and general environment. This is one of the reasons that regulatory agencies typically use additional safety or uncertainty factors when extrapolating laboratory animal data to humans.

Does Research Show That the Compound Affects a Specific Target Organ? Will Humans Be Affected Similarly?

Some toxic agents affect only specific organs and not others. This organ specificity may be due to particular patterns of absorption, distribution, metabolism, and excretion; the presence of specific receptors; or organ function. For example, organ specificity may reflect the presence in the organ of relatively high levels of an enzyme capable of metabolizing or changing a compound to a toxic form of the compound,⁸³ or it may reflect the relatively low level of an enzyme capable of detoxifying a compound. An example of the former is liver toxicity caused by inhaled carbon tetrachloride, which affects the liver but not the lungs because of extensive metabolism to a toxic metabolite within the liver but relatively little such metabolism in the lung.⁸⁴

Some chemicals, however, may cause nonspecific effects or even multiple effects. Lead is an example of a toxic agent that affects many organ systems, including the blood, the central and peripheral nervous systems, the reproductive system, and the kidneys.

83. Certain chemicals act directly to produce toxicity, whereas others require the formation of a toxic metabolite. For example, the potent rat liver carcinogen, N-nitroso-dimethylamine (NDMA, present at trace levels in a typical diet and in some pharmaceutical preparations), requires metabolic activation in order to bind to DNA and cause mutations that lead to cancer in experimental animals. In humans, there appears to be only one enzyme capable of activating NDMA to its reactive intermediate, and this enzyme (called CYP2E1) is only expressed in the liver, suggesting that liver cancer would be the only kind of cancer expected in humans exposed to NDMA. However, other carcinogenic nitrosamines can be activated by enzymes present in other human tissues, as well as the liver. See Yupeng Li & Stephen S. Hecht, *Metabolic Activation and DNA Interactions of Carcinogenic N-Nitrosamines to Which Humans Are Commonly Exposed*, 23 *Int'l J. Molecular Scis.* 4559 (2022), <https://doi.org/10.3390/ijms23094559>.

84. Brian J. Day et al., *Potentiation of Carbon Tetrachloride-Induced Hepatotoxicity and Pneumotoxicity by Pyridine*, 8 *J. Biochem. Toxicology* 11 (1993), <https://doi.org/10.1002/jbt.2570080104>.

Specificity often reflects the function of individual organs. For example, the thyroid is particularly susceptible to radioactive iodine in atomic fallout because thyroid hormone is unique within the body in that it requires iodine. Through evolution, a very efficient and specific mechanism has developed that concentrates any absorbed iodine preferentially within the thyroid, rendering the thyroid particularly at risk from radioactive iodine. In a test tube, the radiation from radioactive iodine can affect the genetic material obtained from any cell in the body, but in the intact laboratory animal or human, only the thyroid is at risk.

The unfolding of the human genome already is beginning to provide information pertinent to understanding the wide variation in human risk from environmental chemicals. The impact of this understanding on toxic tort causation issues continues to be explored.⁸⁵ Several examples of the importance of genetic variability in response to toxic substances is provided later in the section titled “What is the likelihood that the disease would have occurred without the specific exposure?”

What Is Known About the Chemical Structure of the Compound and Its Relationship to Toxicity?

Understanding the structural aspects of chemical toxicology has led to the use of structure activity relationships (SAR) as a formal method of predicting the potential toxicity of new chemicals. This technique compares the chemical structure of compounds with known toxicity and the chemical structure of compounds with unknown toxicity. Toxicity then is estimated based on the molecular similarities between the two compounds. Although SAR is used

85. Nat'l Rsch. Council, *Applications of Toxicogenomic Technologies to Predictive Toxicology and Risk Assessment* (2007), <https://doi.org/10.17226/12037>. Genomics can also be misinterpreted. An example is the use of white blood cell gene expression to determine whether benzene was a cause of acute myelogenous leukemia (AML) in individual workers. Martyn T. Smith, *Misuse of Genomics in Assigning Causation in Relation to Benzene Exposure*, 14 Int'l J. Occupational & Env't Health 144 (2008), <https://doi.org/10.1179/oeh.2008.14.2.144>, describes why the failure to match a pattern of DNA expression in workers with AML who were previously exposed to benzene is not scientifically defensible as a means to establish the lack of causation, as said to have been done in workers' compensation cases in California. The wide range in the rate of metabolism of chemicals is at least partly under genetic control. A study of Chinese workers exposed to benzene found approximately a doubling of risk in people with high levels of either an enzyme that increased the rate of formation of a toxic metabolite or an enzyme that decreased the rate of detoxification of this metabolite. There was a sevenfold increase in risk for those who had both genetically determined variants. Nathan Rothman et al., *Benzene Poisoning, A Risk Factor for Hematological Malignancy, Is Associated with the NQO1 609C→T Mutation and Rapid Fractional Excretion of Chlorzoxazone*, 57 Cancer Rsch. 2839 (1997), <https://perma.cc/4BDM-JXNZ>. See also Lucio G. Costa & David L. Eaton, *Gene Environment Interactions: Fundamentals of Ecogenetics* (2006).

extensively by the EPA in evaluating many new chemicals required to be tested under the registration requirements of the TSCA, its reliability has a number of limitations.⁸⁶

Has the Compound Been the Subject of In Vitro Research, and If So, Can the Findings Be Related to What Occurs In Vivo?

Cellular and tissue culture research can be particularly helpful in identifying mechanisms of toxic action and potential target organ toxicity. The major barrier to the use of in vitro results is the frequent inability to relate doses that cause cellular toxicity to doses that cause whole-animal toxicity. In many critical areas, knowledge that permits such quantitative extrapolation is lacking.⁸⁷ Nevertheless, the ability to quickly test new products through in vitro tests, using

86. For example, benzene and the alkyl benzenes (which include toluene, xylene, and ethyl benzene) share a similar chemical structure. SAR works exceptionally well in predicting the acute central nervous system anesthetic-like effects of both benzene and the alkyl benzenes. Although there are slight differences in dose–response relationships, they are readily explained by the inter-related factors of chemical structure, vapor pressure, and lipid solubility (the brain is highly lipid). Nat’l Rsch. Council, *The Alkyl Benzenes* (1981). However, only benzene produces damage to the bone marrow and leukemia; the alkyl benzenes do not have this effect. This difference is the result of specific toxic metabolic products of benzene in comparison with the alkyl benzenes. Thus, SAR is predictive of neurotoxic effects but not bone marrow effects. See David A. Eastmond et al., *Lymphohematopoietic Cancers Induced by Chemicals and Other Agents and Their Implications for Risk Evaluation: An Overview*, 761 *Mutation Rsch.* 40 (2014), <https://doi.org/10.1016/j.mrrev.2014.04.001>. Advances in computational approaches also show promise in improving SAR. See Nat’l Rsch. Council, *supra* note 63. For an example of how gains in computational sciences and “artificial intelligence” (e.g., “Deep Learning”) are being applied to improve SAR analysis for predicting chemical toxicity, see Gabriel Idakwo et al., *Deep Learning-Based Structure-Activity Relationship Modeling for Multi-Category Toxicity Classification: A Case Study of 10K Tox21 Chemicals With High-Throughput Cell-Based Androgen Receptor Bioassay Data*, 10 *Frontiers Physiology* 2044 (2019), <https://doi.org/10.3389/fphys.2019.01044>. In *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 509 U.S. 579 (1993), the Court rejected a per se exclusion of SAR, animal data, and reanalysis of previously published epidemiological data where there were negative epidemiological data. Most cases involving SAR expert opinion involve patent litigation or analogous drugs under the Controlled Substance Analogue Enforcement Act. See, e.g., *Takeda Pharm. Co. v. Torrent Pharms. Ltd.*, No. 17-3186 (SRC)(CLW), 2020 WL 549594, at *26 (D.N.J. Feb. 4, 2020), *aff’d*, 844 F. App’x 339 (Fed. Cir. 2021); *United States v. Cooper*, No. 3:14-cr-014-J-20MCR, 2015 WL 13850123, at *6, *9 (M.D. Fla. Jan. 14, 2015).

87. See, e.g., *In re Denture Cream Prods. Liab. Litig.*, No. 09-2051-MD, 2015 WL 392021 (S.D. Fla. Jan. 28, 2015) (expert failed to explain relevancy of in vitro studies to humans or account for factors needed to make a proper extrapolation).

human cells, provides invaluable “early warning systems” for toxicity.⁸⁸ For example, screening of new chemicals with a simple bacterial mutagenicity assay such as the Ames test will identify if the new chemical is likely to be mutagenic, and thus carcinogenic. The pharmaceutical industry routinely evaluates new chemicals for mutagenic/carcinogenic potential, and generally will not spend large amounts of time and money in further development if the candidate drug is identified through such simple tests as likely to be mutagenic and thus carcinogenic.

Is the Association Between Exposure and Disease Biologically Plausible?

No matter how strong the temporal relationship between exposure and the development of disease, or the supporting epidemiological evidence, it is difficult to accept an association between a compound and a health effect when no plausible mechanism can be identified by which the chemical or physical exposure leads to the putative effect.⁸⁹

Specific Causal Association Between an Individual’s Exposure and Onset of Disease

An expert who opines that exposure to a compound caused a person’s disease engages in deductive clinical reasoning.⁹⁰ In most instances, cancers and other diseases do not wear labels documenting their causation. The opinion is based on an assessment of the individual’s exposure, including the amount, the temporal relationship between the exposure and disease, and other disease-causing factors. This information is then compared with scientific data on the relationship

88. Despite its limitations, in vitro research can strengthen inferences drawn from whole-animal bioassays and can support opinions regarding whether the association between exposure and disease is biologically plausible. See Klapacz & Gollapudi, *supra* note 61; Rogers, *supra* note 61.

89. See, e.g., *Sarkees v. E.I. DuPont de Nemours & Co.*, No. 17-CV-651, 2020 WL 906331 (W.D.N.Y. Feb. 25, 2020) (expert reviewed animal studies, in vitro studies, and human studies and the inferential steps utilized by researchers to propose a likely mechanism for toxic effects in humans).

90. For an example of deductive clinical reasoning based on known facts about the toxic effects of a chemical and the individual’s pattern of exposure, see Bernard D. Goldstein, *Is Exposure to Benzene a Cause of Human Multiple Myeloma?*, 609 *Annals N.Y. Acad. Scis.* 225 (1990), <https://doi.org/10.1111/j.1749-6632.1990.tb32070.x>.

between exposure and disease. The certainty of the expert's opinion depends on the strength of the research data demonstrating a relationship between exposure and the disease at the dose in question and the presence or absence of other disease-causing factors (also known as confounding factors).⁹¹

Particularly problematic are generalizations made in personal injury litigation from regulatory positions. Regulatory standards are set for purposes far different than determining the preponderance of evidence in a toxic tort case. Further, if regulatory standards are discussed in toxic tort cases to provide a reference point for assessing exposure levels, it must be recognized that there is a great deal of variability in the extent of evidence required to support different regulations.⁹² The extent of evidence required to support regulations depends on

1. the law (e.g., the Clean Air Act National Ambient Air Quality Standard provisions have language focusing regulatory activity for primary pollutants on adverse health consequences to sensitive populations with an adequate margin of safety and with no consideration of economic consequences, while regulatory activity under the TSCA and its revision, the Lautenberg Act, clearly asks for some balance between the societal benefits and risks of new chemicals⁹³);
2. the specific endpoint of concern (e.g., consider the concern caused by cancer and adverse reproductive outcomes versus almost anything else); and

91. Causation issues are discussed in Steve C. Gold et al., *Reference Guide on Epidemiology*, section titled "General Causation and Specific Causation," and John B. Wong et al., *Reference Guide on Medical Testimony*, section titled "Medical Decision-Making," both in this manual. For a detailed analysis of the challenges of assessing epidemiological and toxicological data to establish causal relationships between exposure and disease, see Nat'l Acads. Scis., Eng'g, & Med., *Advancing the Framework for Assessing Causality of Health and Welfare Effects to Inform National Ambient Air Quality Standard Reviews* (2022).

92. See, e.g., *Kirk v. Schaeffler Grp. USA, Inc.*, 887 F.3d 376, 392 (8th Cir. 2018) (noting that regulatory standards are "designed to protect public health," not to be a measure of causation, and that regulatory agencies may set standards "without having precise data on the question of how much harm, or what kind of harm, some specific amount of . . . substance might reasonably be expected to cause" (quoting *Wright v. Willamette Indus., Inc.*, 91 F.3d 1105, 1107 (8th Cir. 1996)); *Williams v. Mosaic Fertilizer, LLC*, 889 F.3d 1239, 1247 (11th Cir. 2018) (stating that regulatory standards are designed to be protective whereas "dose-response calculations aim to identify the exposure levels that actually cause harm").

93. *Cnty. Voice v. U.S. Env't Prot. Agency*, 997 F.3d 983, 990–91 (9th Cir. 2021) (citing *Whitman v. Am. Trucking Ass'ns*, 531 U.S. 457, 467–68 (2001)) (explaining TSCA's consideration of "non-health factors such as achievability and cost" and the opposite for the Clean Air Act, which the Supreme Court has held "does not" permit "the EPA to consider costs in setting clean air standards"). See also *Nat'l Res. Defense Council v. U.S. Env't Prot. Agency*, 38 F.4th 34 (9th Cir. 2022) (challenge to the EPA's conclusions on human health risks and cost-benefit analysis for glyphosate under FIFRA, based on the EPA's failure to follow its Cancer Guidelines, resulting in court vacating human health portion of the EPA's Interim Decision and remanding).

3. the societal impact (e.g., the public's support for control of an industry that causes air pollution versus the public's relative lack of desire to alter personal automobile use patterns).

These three concerns, as well as others, including costs, politics, and the virtual certainty of litigation challenging the regulation, have an impact on the level of scientific proof required by the regulatory decision-maker.⁹⁴

In addition, regulatory standards traditionally include protective factors to reasonably ensure that susceptible individuals are not put at significant risk. Furthermore, standards often are based on the risk that results from lifetime exposure. Accordingly, the mere fact that an individual has been exposed to a level above a standard does not necessarily mean that an adverse effect has occurred or will occur.

Was the Plaintiff Exposed to the Substance, and If So, Did the Exposure Occur in a Manner That Can Result in Absorption into the Body?

Evidence of exposure is essential in determining the effects of harmful substances.⁹⁵ Basically, potential human exposure is measured in one of three ways. First, when direct measurements cannot be made, exposure can be measured by

94. These concerns are discussed in Stephen Breyer, *Breaking the Vicious Circle: Toward Effective Risk Regulation* (1993).

95. “[S]cientific knowledge of the harmful level of exposure to a chemical” is a ‘minimal fact’ necessary for establishing causation.” *Cooper v. Meritor, Inc.*, No. 4:16-CV-52-DMB-JMV, 2019 WL 545271, at *5 (N.D. Miss. Feb. 11, 2019) (quoting *Allen v. Pa. Eng’g Corp.*, 102 F.3d 194, 199 (5th Cir. 1996)). In toxic tort cases, the plaintiff bears the burden of demonstrating the level of exposure. *Zellers v. NexTech Ne. LLC*, 895 F. Supp. 2d 734, 741 (E.D. Va. 2012) (citing *Westberry v. Gislaved Gummi AB*, 178 F.3d 257, 260 (4th Cir. 1999)). The court in *Zellers* discussed cases from the Fourth, Fifth, and Seventh Circuit U.S. Courts of Appeal where expert opinions were or were not found to be reliable, noting that where “substantial exposure” is established by the expert, “more detailed quantitative data about [the plaintiff’s] level of exposure” may not be necessary. 895 F. Supp. 2d at 739 (citing *Westberry*, 178 F.3d at 264). On the other hand, where there is “no direct evidence of the plaintiff’s level of exposure to the chemical at issue,” the expert’s opinion may be excluded as unreliable. *Zellers*, 895 F. Supp. 2d at 740 (citing *Allen*, 102 F.3d at 199). Likewise, where the expert did not review medical records, perform any tests or calculations, or seek to learn any specifics about the environment in which the alleged exposure took place, the expert’s opinion as to causation based on exposure level was excluded as being “not based on sufficient information about the level of the plaintiff’s exposure to the alleged toxin.” 895 F. Supp. 2d at 740 (citing *Wintz v. Northrop Corp.*, 110 F.3d 508, 510–13 (7th Cir. 1997)). In *Zellers*, the expert opinions were ultimately excluded because the experts “did not consider the extent of Plaintiffs’ exposure,” rendering their opinions “speculative at best.” 895 F. Supp. 2d at 741, *aff’d sub nom.* *Zellers v. NexTech Ne., LLC*, 533 F. App’x 192, 198 (4th Cir. 2013). For a discussion of general issues

mathematical modeling, in which one uses a variety of physical factors to estimate the transport of the pollutant from the source to the individual. For example, mathematical models take into account such factors as wind variations to allow calculation of the transport of radioactive iodine from a federal atomic research facility to nearby residential areas. Second, exposure can be directly measured in the medium in question—air, water, food, or soil. When the medium of exposure is water, soil, or air, hydrologists or meteorologists may be called upon to contribute their expertise to measuring exposure. The third approach directly measures human individuals through some form of biological monitoring, such as blood tests to determine blood lead levels or urinalyses to check for a urinary metabolite indicative of pollutant exposure. Ideally, both environmental testing and biological monitoring are performed; however, this is not always possible, particularly in instances of past exposure (unless the chemical(s) had very long half-lives (e.g., greater than three or four years), in which case blood levels may be useful to demonstrate significant exposures that occurred many years earlier).⁹⁶

The toxicologist must go beyond understanding exposure to determine if the individual was exposed to the compound in a manner that can result in absorption into the body. The absorption of the compound is a function of its physiochemical properties, its concentration, and the presence of other agents or conditions that assist or interfere with its uptake. For example, inhaled lead is absorbed almost totally, whereas ingested lead is taken up only partially into the body. Of toxicological relevance is the finding that children absorb ingested lead substantially more efficiently than adults, which may contribute to the enhanced susceptibility of children to the neurotoxic effects of lead.⁹⁷ Iron deficiency and low nutritional calcium intake, both common conditions of children living in impoverished urban environments, increase the amount of ingested lead that is absorbed in the gastrointestinal tract and passes into the bloodstream.⁹⁸

regarding assessment of exposure of toxic substances see M. Elizabeth Marder & Joseph V. Rodricks, *Reference Guide on Exposure Science and Exposure Assessment*, in this manual.

96. See sections titled “Toxicology and the Law” and “Use of Biomarkers in Toxicology Exposure Assessment” above.

97. Alan R. Abelsohn & Margaret Sanborn, *Lead and Children: Clinical Management for Family Physicians*, 56 Canadian Fam. Physician 531 (2010), <https://perma.cc/M6S4-P2L7>.

98. The term *bioavailability* is used to describe the extent to which a compound, such as lead, is taken up into the body. In essence, bioavailability is at the interface between exposure and absorption into the organism. See, e.g., *In re Denture Cream Prods. Liab. Litig.*, No. 09-2051-MD, 2015 WL 392021 (S.D. Fla. Jan. 28, 2015) (expert’s analysis of in vitro studies failed to demonstrate bioavailability of zinc in denture cream); *Ridings v. Maurice*, No. 15-00020-CV-W-JTM, 2019 WL 13159921, at *4 (W.D. Mo. Aug. 12, 2019) (permitting expert to testify regarding bioavailability of drug alleged to have caused, at least in part, the plaintiff’s injuries). Under certain circumstances bioavailability is not a necessary condition to establish general causation, however. *Adkisson v. Jacobs Eng’g Grp., Inc.*, 342 F. Supp. 3d 791, 805–07 (E.D. Tenn. 2018) (“Biological plausibility and bioavailability are important scientific concepts. But it does not appear that either is strictly

Were Other Factors Present That Can Affect the Distribution of the Compound Within the Body?

Once a compound is absorbed into the body through the skin, lungs, or gastrointestinal tract, it is distributed throughout the body via the bloodstream. Thus, the rate of distribution depends on the rate of blood flow to various organs and tissues. Distribution and resulting toxicity also are influenced by other factors, including the dose, the route of entry, tissue solubility, lymphatic supplies to the organ, metabolism, and the presence of specific receptors or uptake mechanisms within body tissues.

What Is Known About How Metabolism in the Human Body Alters the Toxic Effects of the Compound?

Metabolism is the alteration of a chemical by bodily processes. It does not necessarily result in less toxic compounds being formed. In fact, many of the organic chemicals that are known human cancer-causing agents require metabolic transformation before they can cause cancer. A distinction often is made between direct-acting agents, which cause toxicity without any metabolic conversion, and indirect-acting agents, which require metabolic activation before they can produce adverse effects. Metabolism is complex, because a variety of pathways compete for the same agent; some produce harmless metabolites, and others produce toxic agents.⁹⁹

What Excretory Route Does the Compound Take, and How Does This Affect Its Toxicity?

Excretory routes are urine, feces, sweat, saliva, expired air, and lactation. Many inhaled volatile agents are eliminated primarily by exhalation. Small water-soluble compounds are usually excreted through urine. Higher-molecular-weight

necessary for an association between a particular toxic agent and a particular disease to be considered causal. Accordingly, neither is required to establish proof of general causation.”).

99. Courts have explored the relationship between metabolic transformation and carcinogenesis. See, e.g., *In re E.I. Du Pont De Nemours & Co. C-8 Personal Inj. Litig.*, No. 2:13-md-2433, 2016 WL 3064124, at *4 (S.D. Ohio May 30, 2016) (“For example, one pound of a toxic, cancer-causing chemical that biopersist in humans for decades and bioaccumulates in the environment for millennia may cause more harm than 100 pounds of a toxic, cancer-causing chemical that metabolizes and/or degrades quickly.”).

compounds are often excreted through the biliary tract into the feces. Certain fat-soluble, poorly metabolized compounds, such as PCBs, may persist in the body for decades, although they can be excreted in the milk fat of lactating women, with potential effects in their infants.

Does the Temporal Relationship Between Exposure and the Onset of Disease Support or Contradict Causation?

In acute toxicity, there is usually a short time period between cause and effect. However, in some situations, the length of basic biological processes necessitates a longer period of time between initial exposure and the onset of observable disease. For example, in AML (acute myelogenous leukemia), the adult form of acute leukemia, at least one to two years must elapse from initial exposure to radiation, benzene, or cancer chemotherapy before the manifestation of a clinically recognizable case of leukemia, and the period of significantly higher risk from the last exposure usually persists for no more than about fifteen years. A toxic tort claim alleging a shorter or longer time period between cause and effect is scientifically highly debatable. Much longer latency periods are necessary for the manifestation of solid tumors caused by agents such as asbestos and arsenic.¹⁰⁰

100. The temporal relationship between exposure and causation is discussed in many cases. See, e.g., *Johnson v. Arkema, Inc.*, 685 F.3d 452, 466–67 (5th Cir. 2012) (quoting *Curtis v. M&S Petroleum, Inc.*, 174 F.3d 661, 670 (5th Cir. 1999)) (“[T]emporal connection standing alone is entitled to little weight in determining causation[.]” but may be “entitled to greater weight when there is an established scientific connection between exposure and illness or other circumstantial evidence supporting the causal link.”) (internal citations omitted); *In re Paulsboro Derailment Cases*, 746 F. App’x 94, 99 (3d Cir. 2018) (citing *Kannankeril v. Terminix Int’l, Inc.*, 128 F.3d 802, 808–09 (3d Cir. 1997)) (noting that an expert may only rely on a temporal relationship analysis if the expert uses that relationship within the greater methodology of a differential diagnosis); *C.W. ex rel. Wood v. Textron, Inc.*, 807 F.3d 827, 838 (7th Cir. 2015) (quoting *Ervin v. Johnson & Johnson*, 492 F.3d 901, 904–05 (7th Cir. 2007)) (rejecting expert methodology relying on the relationship between the time of exposure and the onset of injury, explaining that “[t]he mere existence of a temporal relationship between taking a medication and the onset of symptoms does not show a sufficient causal relationship”); *Arias v. DynCorp*, 928 F. Supp. 2d 10, 8 (D.D.C. 2013), *aff’d*, 752 F.3d 1011 (D.C. Cir. 2014) (discussing the relationship between temporal evidence and specific causation); *Leake v. United States*, 843 F. Supp. 2d 554, 561–62 (E.D. Pa. 2011) (noting that “even a strong temporal relationship between exposure and injury, in and of itself,” may not be “sufficient to establish a reliable general causation opinion”).

When Exposure to the Substance Is Associated with the Disease, Is There a Benchmark Dose or No Observed Adverse Effect Level, and If So, Was the Individual Exposed Above This Level?

As discussed in the section titled “In Vivo Research: Use of Live Animals in Toxicity Testing and Safety Assessment” above, for agents that produce effects other than through mutations, it is assumed that there is some level below which no toxic effects will occur—the so-called threshold. Many chemicals may have multiple different toxic effects, and the dose–response relationship for each adverse outcome (toxic effect) may be different. Typically, regulatory agencies use the most sensitive toxic effect (the effect associated with the lowest dose, or threshold). In toxic tort litigation related to a specific effect (e.g., a particular type of birth defect, liver or kidney disease, etc.), experts may argue that the dose of the agent was below the threshold and thus insufficient to have caused or substantially contributed to the disease. If the level of exposure was below this level, a relationship between the exposure and disease cannot be established.¹⁰¹ When only laboratory animal data are available, the expert extrapolates the BMD, NOAEL, or other metric of a threshold level based on available experimental data. Occasionally, the lowest dose used has an observable effect, and thus no NOAEL can be established. In those circumstances, the lowest dose used is called the Lowest Observed Adverse Effect Level (LOAEL). As discussed in the section noted above, in regulatory toxicology, if using a NOAEL value to estimate a threshold dose, the expert decreases this level by one or more safety factors (usually a factor of ten for each one) to ensure no human effect.¹⁰² As discussed in the section titled “Benchmark Dose (BMD),” if using a BMD, the lower statistical confidence level (usually tenth percentile) of the BMD (called

101. See, e.g., *Wyman v. U.S. Surgical Corp.*, No. 1:18-cv-00095-JAW, 2020 WL 1932338, at *21 (D. Me. Apr. 21, 2020) (In the context of mercury, benchmark dose is “the amount . . . associated with the threshold for a health effect on sensitive populations according to epidemiological studies.”); *Kolakowski v. Sec’y of Health & Hum. Servs.*, No. 99-0625V, 2010 WL 5672753, at *11 (Fed. Cl. Nov. 23, 2010) (Benchmark dose “is a dose at which some statistical interval, measured from a percentage of the population with a physical response to the dose.”); *In re Valsartan, Losartan, and Irbesartan Prods. Liab. Litig.*, No. 2022 WL 807343, at *2 (D.N.J. Mar. 17, 2022) (permitting an expert to testify using benchmark-dose methodology, acknowledging that while “this is not the methodology used by many scientists,” it “is nonetheless used by some in the relevant scientific community”).

102. See, e.g., *supra* note 54 & accompanying text; Robert G. Tardiff & Joseph V. Rodricks, *Toxic Substances and Human Risk: Principles of Data Interpretation* 391 (1987); Joseph V. Rodricks, *Calculated Risks* 230–39 (2d ed. 2006). For regulatory toxicology, NOAEL is being replaced by a more statistically robust approach known as the benchmark dose. See *supra* note 55 & accompanying text. For example, the EPA’s use of the benchmark dose takes into account comprehensive dose–response information, unlike NOAEL.

the BMDLow, or BMDL, or BMDL₁₀) can also be calculated from animal data or from human toxicity data if they exist. The estimation of a threshold dose using either NOAEL or BMD data, however, is generally not applied to substances that exert toxicity by causing DNA damage and mutations that may lead to cancer. Theoretically, any exposure at all to mutagens may increase the risk of cancer, although the risk may be very small and not achieve statistical significance.¹⁰³

Although the concept of the non-threshold response is generally used only for genotoxic carcinogens or other outcomes associated with gene mutations (e.g., development of some types of birth defects), it is sometimes claimed that other toxic responses, such as the effects of lead on children's IQ, also follow a linear response at low doses. However, it is biologically likely that there are thresholds for such responses, but if the threshold is below background levels of exposure (such as may be the case with lead), the dose–response would appear to be linear since a threshold dose could not be established and background exposures exceed the putative threshold level.

What Is the Likelihood That the Disease Would Have Occurred Without the Specific Exposure?

In coming to an expert opinion as to whether the exposure to a known cause of a disease is likely to be responsible for the case observed in the plaintiff, the expert must consider other potential causes of the same disease. Some may be

103. See sources cited *supra* notes 57 and 65. See also *Myers v. United States*, No. 02cv1349–BEN, 2014 WL 6611398, at *39 (S.D. Cal. Nov. 20, 2014) (explaining that although thallium may be harmful at certain levels, “[s]ince we are exposed to thallium on a daily basis, there must be a level that does not cause adverse health effects”). U.S. regulatory approaches aimed at protecting the general population tend to avoid setting a standard for a known human carcinogen, because any allowable level below the standard is at least theoretically capable of causing cancer. However, exposure to many chemical carcinogens, including benzene and arsenic, cannot be eliminated. Thus, agencies and Congress have developed a number of ingenious means to regulate carcinogens while not seeming to acquiesce in exposure of the general population to a carcinogen. These include the FDA's approach to de minimis risk and the EPA's setting of a zero maximum contaminant level goal for carcinogens in drinking water while setting a maximum contaminant level above zero that is set as closely as possible to the MCLG, taking technology and cost data into account. In contrast, occupational standards, which also take into account feasibility, permit exposure to known human carcinogens. A generally outmoded approach for environmental or indoor air guidelines has been to divide the permissible OSHA standard by a factor accounting for the presumed lifetime exposure to the environmental chemical compared with forty-five years at a forty-hour workweek. Plaintiff's experts in a toxic court case must rely on more than exposure levels established by regulatory agencies, as those levels “often build in considerable cushion to account for the most sensitive members of the population.” *Williams v. Mosaic Fertilizer, LLC*, 889 F.3d 1239, 1246–47 (11th Cir. 2018) (affirming exclusion of expert opinion for, inter alia, relying on regulatory standards of exposure levels).

due to other known causal factors, such as lung cancer in a uranium miner who is also a cigarette smoker. In addition, with rare exceptions, almost all diseases have some anticipated background causes that appear related to human biology, particularly aging. An additional challenge comes in unraveling the contribution of a specific substance when exposure may have been to multiple agents. There are numerous examples where the effect of two combined exposures to two substances both known to cause the same disease is greater than the sum of the individual effects (synergism). The example above of lung cancer in uranium miners who also smoke is a good illustration. Another example of such synergism is the incidence of liver cancer in people who are exposed to the dietary contaminant aflatoxin B1, and also are infected with hepatitis B virus. Both are known causes of liver cancer (relative risk for hepatitis B virus antigen positivity is approximately tenfold; relative risk for aflatoxin B1 alone is approximately threefold) but the presence of both risk factors increased the risk of developing liver cancer by approximately sixty times.¹⁰⁴

Not surprisingly, cancer has received the most attention in research aimed at disentangling external and internal causes. For some cancers, such as lung cancer, smoking is an obvious predominant external cause, and the extent to which other known causes contribute to overall lung cancer incidence also can be estimated. But these known external causes do not currently account for all cases of lung cancer. For other cancers, such as pancreatic cancers, it has been very difficult to find any external cause. There are, however, a few instances in which the exposure to a specific chemical is nearly uniquely associated with a particular (usually rare) type of cancer. For example, the industrial chemical vinyl chloride, widely used in the synthesis of certain plastics, is associated with a rare form of cancer of the blood vessels in the liver (angiosarcoma).¹⁰⁵ Likewise, certain types of mesothelioma (a cancer of the lining of the chest cavity or stomach/intestine) are associated with prior exposure to asbestos, especially among smokers, suggesting a synergistic interaction between smoking and asbestos exposure.¹⁰⁶

That mutations in the DNA of stem cells of virtually every tissue accumulate with age is now generally recognized. Upwards of thousands of mutations

104. John D. Groopman et al., *Aflatoxin and Hepatitis B Virus Biomarkers: a Paradigm for Complex Environmental Exposures and Cancer Risk*, 1 *Cancer Biomarkers* 5 (2005), <https://doi.org/10.3233/cbm-2005-1103>. See also Joshua Jin et al., *Synergism in Actions of HBV with Aflatoxin in Cancer Development*, 499 *Toxicology* 153652 (2023), <https://doi.org/10.1016/j.tox.2023.153652>.

105. The relative risk (RR) of angiosarcoma of the liver was ninety-seven in one of the first occupational cohort studies to demonstrate the association between occupational exposure to vinyl chloride and this very rare form of liver cancer. There is also suggestive evidence that vinyl chloride can contribute to the more common form of liver cancer (hepatocellular carcinoma), although the relative risks are much lower. See Ugo Fedeli et al., *Occupational Exposure to Vinyl Chloride and Liver Diseases*, 25 *World J. Gastroenterology* 4885 (2019), <https://doi.org/10.3748/wjg.v25.i33.4885>.

106. Sonja Klebe et al., *Asbestos, Smoking and Lung Cancer: An Update*, 17 *Int'l J. Env't Rsch. & Pub. Health* 258 (2019), <https://doi.org/10.3390/ijerph17010258>.

per cell are not uncommon in the cells of the elderly. Certain of these mutations, often called driver mutations, are more commonly associated with cancer.

A major cause for the accumulation of mutations as we age is related to the frequency of mistakes made in copying DNA. These reflect the fallibility of normal cell DNA replication as well as the increasing failure with age of DNA repair enzymes to successfully carry out appropriate copy editing of the replicated DNA.

There is increasing evidence that many cancers are associated with heritable genetic differences among individuals. For example, women who inherit one mutated copy of the so-called “breast cancer genes,” *BRCA1* and *BRCA2*, are at substantially greater risk of developing breast cancer when compared to women who have the common (“wild type”) genetic form.¹⁰⁷ The study of gene–environment interactions is of growing importance in the field of toxicology as well as medicine, as there is now substantial evidence that many diseases result from a combination of certain genetic variants and exposure to specific chemicals.¹⁰⁸ For many cancers, genetic variants that decrease the efficiency and accuracy of DNA repair are widely recognized as “susceptibility genes.” The evidence suggests that individuals with genetic deficiency in a DNA repair gene are likely to be more susceptible to the mutagenic effects of chemicals that cause the specific type of mutation repaired by the deficient DNA repair gene.

Estimates of what percent of individual cancer types are due to external versus internal factors are beginning to appear and can be expected to improve with advances in stem cell biology. (Unfortunately, cancers due to intrinsic factors, such as the failure of DNA repair, have been described as “bad luck” to distinguish them from cancers due to external factors.) This information will likely become increasingly important in considering the differential causation of individual cancer types.¹⁰⁹ It is likely that the introduction of potential genetic

107. Women who have mutated copies of BRCA genes have a lifetime risk of developing breast cancer of 45–75%, compared with a “background” lifetime risk of ~10–15%. About 5–10% of all breast cancers can be attributed to genetic risk. See Zora Baretta et al., *Effect of BRCA Germ-line Mutations on Breast Cancer Prognosis: A Systematic Review and Meta-Analysis*, 95 *Medicine* e4975 (2016), <https://doi.org/10.1097/MD.00000000000004975>.

108. See Nat’l Inst. Env’t Health Scis., *Gene and Environment Interaction*, <https://perma.cc/X5HB-39TV>, for a general discussion of how genes and the environment can interact to produce certain diseases.

109. A readily understandable account of the chemistry and biological implications of DNA repair is in the Popular Science Background piece accompanying the award of the 2015 Nobel Prize for Chemistry, DNA Repair—Providing Chemical Stability for Life, <https://perma.cc/S9VL-9KRH>. See also *McManaway v. KBR, Inc.*, No. H-10-1044, 2012 WL 13059744 (S.D. Tex. Aug. 22, 2012) (expert failed to take into account whether alleged genetic transformation injuries caused by sodium dichromate were repaired, rendering his opinion on persistence of these injuries unreliable). A brief overview of the literature on the extent of estimated external and internal causes in different human cancer types is in Bernard D. Goldstein & Varun Patel, *Controversy About the “Bad Luck” Cancer Hypothesis Could Lead to a Useful Tool for Planning Primary Prevention Cancer*

susceptibility factors to environmental/occupational exposures to certain chemicals will become more common in toxic tort litigation, potentially used by plaintiff and defense lawyers alike to address the likelihood of a causal association between an exposure and a disease.

A relatively new and important area of genetic research, called epigenetics, has identified that there can be heritable changes (passed on from one generation to the next) in the functions of DNA that are not associated with a change in the DNA sequence (i.e., not the result of a mutation). Epigenetics (literally translated as “above genetics”) describes how the expression of DNA is regulated by several different biological processes.¹¹⁰ Importantly, there are a growing number of examples where chemicals found in the diet, workplace, and general environment can cause epigenetic changes that may contribute to a wide variety of adverse effects, including many types of cancers, diabetes and metabolic syndrome, neurological and cognitive dysfunction, as well as alterations in the functions of important organs such as the lung, the heart, the immune system, and reproductive organs. Indeed, a search of the scientific literature today shows over 127,000 scientific papers published in the general field of epigenetics. Although much of this scientific literature is focused on the normal (endogenous) functioning of epigenetic phenomena in chronic disease processes, there is growing recognition that many environmental and dietary factors can alter epigenetic processes, including some heavy metals, certain pesticides, some chemicals used in plastics, diesel exhaust, tobacco

Research, 32 Chem. Rsch. Toxicology, 949 (2019), <https://doi.org/10.1021/acs.chemrestox.8b00390>. For a discussion of methodologies for ruling out idiopathic (unknown) causes of a disease, see *Hardeman v. Monsanto Co.*, 997 F.3d 941, 953 (9th Cir. 2021) (quoting *Wendell v. GlaxoSmithKline LLC*, 858 F.3d 1227, 1233–34 (9th Cir. 2017)) (ruling out idiopathy for disease with 70% idiopathy rate where expert relied on clinical experience, literature, and medical records).

110. For example, many of the nucleotide bases that are strung together to make up DNA can have methyl ($-\text{CH}_3$) groups attached to them in specific positions by enzymes known as DNA methyl transferases. DNA methylation functions somewhat like a switch, providing critical information on when a gene is “turned on” or “turned off.” In a similar manner, certain enzymes can modify proteins called chromatin that are intimately associated with DNA and help to regulate when specific genes are turned on or off (gene expression). One group of these chromatin proteins, called histones, can be modified by attaching, or detaching, acetyl groups on the histone proteins (histone acetylases and histone deacetylases), which then modulate whether the gene is expressed or not. This epigenetic process is known as histone modification (histone acetylation or deacetylation). Several other biological processes that also modify the expression of genes via epigenetic processes include protein phosphorylation, ubiquitination, and sumoylation of chromatin proteins. Similar to histone acetylation, adding these different chemical entities to proteins that are involved in regulation of DNA function can modify how genes are expressed, without changing the actual sequence of the DNA, and thus are referred to as epigenetic effects. Other forms of RNA called short non-coding RNA and microRNAs have been found to be important epigenetic regulators of expression of DNA. See Lian Zhang et al., *Epigenetics in Health and Disease*, in *Epigenetics in Allergy and Autoimmunity* (Christopher Chang & Qianjun Lu eds., 2020), https://doi.org/10.1007/978-981-15-3449-2_1.

smoke, polycyclic aromatic hydrocarbons, hormones, radioactivity, viruses and bacteria, to name a few.¹¹¹

As the role of epigenetic factors in cancer and in other diseases unfolds, these might also be grist for the mill of both plaintiff and defense lawyers.

Medical History

Is the Medical History of the Individual Consistent with the Toxicologist's Expert Opinion Concerning the Injury?

One of the basic and most useful tools in diagnosis and treatment of disease is the patient's medical history.¹¹² A thorough, standardized patient information questionnaire would be particularly useful for identifying the etiology, or causation, of illnesses related to toxic exposures; however, there is currently no validated or widely used questionnaire that gathers all pertinent information.¹¹³ Nevertheless, it is broadly recognized that a thorough medical history involves the questioning and examination of the patient as well as appropriate medical testing. The patient's written medical records also should be examined.

The following information is relevant to a patient's medical history: past and present occupational and environmental history and exposure to toxic agents; lifestyle characteristics (e.g., use of nicotine and alcohol); family medical history (i.e., medical conditions and diseases of relatives); and personal medical history (i.e., present symptoms and results of medical tests as well as past injuries, medical conditions, diseases, surgical procedures, and medical test results).

111. See Bob Weinhold, *Epigenetics: The Science of Change*, 114 *Env't Health Perspectives* A160 (2006), <https://doi.org/10.1289/ehp.114-a160>; Giacomo Cavalli & Edith Heard, *Advances in Epigenetics Link Genetics to the Environment and Disease*, 571 *Nature* 489 (2019), <https://doi.org/10.1038/s41586-019-1411-0>.

112. For a thorough discussion of the methods of clinical diagnosis, see John B. Wong et al., *Reference Guide on Medical Testimony*, in this manual. A number of cases have considered the admissibility of the treating physician's opinion based, in part, on medical history, symptomatology, and laboratory and pathology studies. See, e.g., *Morrow v. Brenntag Mid-South, Inc.*, 505 F. Supp. 3d 1287, 1290 (M.D. Fla. 2020) (declining to find expert testimony reliable where the expert, a physician, had not reviewed any "prior medical history or treatment records"); cf. *Galvez v. KLLM Transp. Servs., LLC*, 575 F. Supp. 3d 748, 760–61 (N.D. Tex. 2021) (finding expert testimony reliable where causation opinion was supported by a review of the patient's medical history, part of "the well-established 'scientific method' of diagnosis in the medical community").

113. Office of Tech. Assessment, U.S. Congress, *Reproductive Health Hazards in the Workplace* 8 (1985), at 365–89.

In some instances, the reporting of symptoms in itself can be diagnostic of exposure to a specific substance, particularly in evaluating acute effects.¹¹⁴ For example, individuals acutely exposed to organophosphorus pesticides report headaches, nausea, and dizziness accompanied by anxiety and restlessness. Other reported symptoms are muscle twitching, weakness, and hypersecretion with sweating, salivation, and tearing.¹¹⁵

Are the Complaints Specific or Nonspecific?

Acute exposure to many toxic agents produces a constellation of nonspecific symptoms, such as headaches, nausea, lightheadedness, and fatigue. These types of symptoms are part of human experience and can be triggered by a host of medical and psychological conditions. They are almost impossible to quantify or document beyond the patient's report. Thus, these symptoms can be attributed mistakenly to an exposure to a toxic agent or discounted as unimportant, when in fact they reflect a significant exposure.¹¹⁶

114. See *McManaway v. KBR, Inc.*, No. H-10-1044, 2012 WL 13059744, at *6 (S.D. Tex. Aug. 22, 2012) (expert review of medical records and in-person interviews, including review of acute symptoms, supports a finding of reliability); *Coene v. 3M Co.*, No. 10-CV-6546-FPG, 2017 WL 1046749, at *3, *9 (W.D.N.Y. Mar. 20, 2017) (finding toxicologist's opinion regarding causation of silicosis to be reliable when the opinion was based on a review of the plaintiff's medical records as well as conversations with the plaintiff, the MSDS sheets for the chemicals the plaintiff worked with that allegedly caused the injury, and technical literature); *Ledbetter v. Blair Corp.*, No. 3:09-CV-843-WKW, 2012 WL 2464000 (M.D. Ala. June 27, 2012) (finding medical toxicologist's opinion as to impairment caused by certain prescription drugs to be reliable when it was based, in part, on a review of medical and prescription records). *Contra* *Gilbert v. Lands' End Inc.*, No. 19-cv-823-jdp, 2022 WL 2643514, at *12-13 (W.D. Wis. July 8, 2022) (in a case alleging injury from required work uniforms, finding expert's opinion unreliable when it was based only on a review of partial medical records and when the expert did not talk to the plaintiffs, "attempt to link any particular plaintiffs' symptom to any particular garment," or "rule out other potential causes").

115. EPA, *Recognition and Management of Pesticide Poisonings* 45 (6th ed. 2013), <https://perma.cc/Z4UL-9NCQ>.

116. Complex issues of exposure to oil and dispersants, chemicals, fumes, and odors and reported adverse health effects are addressed in a medical settlement establishing a Periodic Medical Consultation Program and a Specified Physical Conditions Matrix as discussed in *In re Oil Spill by Oil Rig Deepwater Horizon*, 295 F.R.D. 112 (E.D. La. 2013) (medical class certified for settlement purposes only). An example of a disease that may be caused by exposure to toxins but may present only with nonspecific symptoms is Legionnaires' Disease, explored at length in *Mueller v. Chugach Federal Solutions, Inc.*, No. 12-S-00624-NE, 2014 WL 2891030 (N.D. Ala. June 25, 2014). There, the decedent's wife alleged that her husband had been exposed to *Legionella pneumophila* at his workplace. *Id.* at *1. The court noted that the "disease may present early in the illness with non-specific symptoms, so it can be difficult to diagnose." *Id.* at *20. Indeed, the decedent in *Mueller* had been hospitalized for pneumonia and treated for the same prior to finally being diagnosed with Legionnaires' Disease after a urine antigen test. *Id.* at *4. In opining that that decedent's illness was

In taking a careful medical history, the expert focuses on the time pattern of symptoms and disease manifestations in relation to any exposure and on the constellation of symptoms to determine causation. It is easier to establish causation when symptoms or a specific diagnosis are unusual and rarely caused by anything other than the suspect chemical (e.g., blue lips and shortness of breath caused by dyes that produce methemoglobinemia) or rare cancers such as hemangiosarcoma, associated with vinyl chloride exposure, and mesothelioma, associated with asbestos exposure. However, many cancers and other conditions are associated with several extrinsic or intrinsic causative factors, complicating proof of causation.¹¹⁷

Do Laboratory Tests Indicate Exposure to the Compound?

Two types of laboratory tests can be considered: tests that are routinely used in medicine to detect changes in normal body status, and specialized tests, which are used to directly or indirectly detect the presence of the chemical or physical agent.¹¹⁸ For the most part, tests used to demonstrate the presence of a toxic agent are frequently unavailable from clinical laboratories. Even when available from a hospital or a clinical laboratory, a test such as that for carbon monoxide bonded to hemoglobin (carboxyhemoglobin) is done so rarely that there may be concerns regarding its accuracy. Other tests, such as the test for blood lead levels, are required for routine surveillance of potentially exposed workers.¹¹⁹

caused by exposure in the workplace, the expert was permitted to base his opinions on various materials produced during litigation, his own experience, and the fact that other employees also exhibited symptoms consistent with Legionnaires' Disease. *Id.* at *6.

117. Failure to rule out other potential causes of symptoms may lead to a ruling that the expert's report is inadmissible. *See, e.g.,* *Hardeman v. Monsanto Co.*, 997 F.3d 941, 953, 965 (9th Cir. 2021) (expert's specific causation opinion was admissible where hepatitis C was ruled out as a cause of plaintiff's non-Hodgkins lymphoma based on research studies and a temporal analysis of exposures and active disease); *Pluck v. BP Oil Pipeline Co.*, 640 F.3d 671, 678–81 (6th Cir. 2011) (affirming district court's conclusion that expert's opinion was unreliable when the expert failed to rule out alternative causes of the plaintiff's non-Hodgkins lymphoma); *Scott v. Dyno Nobel, Inc.*, No. 4:16-CV-1440 HEA, 2021 WL 1750238, at *10–11 (E.D. Mo. May 4, 2021) (finding expert opinion reliable when it ruled out other possible causes of symptoms and noting that another expert's opinion was found inadmissible for failure to do the same).

118. *See, e.g.,* *Myers v. United States*, No. 02cv1349–BEN, 2014 WL 6611398, at *35–38 (S.D. Cal. Nov. 20, 2014) (explaining at length the necessity of reliable laboratory testing and reasons why tests may be found unreliable).

119. However, if a laboratory is certified for the testing of blood lead levels in workers, for which the OSHA action level is 40 micrograms per deciliter (µg/dl), it does not necessarily mean that it will give reliable data on blood lead levels at the much lower CDC Reference Level of 3.5

What Other Causes Could Lead to the Given Complaint?

With few exceptions, acute and chronic diseases, including cancer, can be caused by either a single toxic agent or a combination of agents or conditions. In taking a careful medical history, the expert examines the possibility of competing causes, or confounding factors, for any disease, which leads to a differential diagnosis. In addition, ascribing causality to a specific source of a chemical requires that a history be taken concerning other sources of the same chemical. The failure of a physician to elicit such a history or of a toxicologist to pay attention to such a history raises questions about competence and leaves open the possibility of competing causes of the disease.¹²⁰

Is There Evidence of Interaction with Other Chemicals?

An individual's simultaneous exposure to more than one chemical may result in a response that differs from that which would be expected from exposure to only one of the chemicals.¹²¹ When the effect of multiple agents is that which would be predicted by the sum of the effects of individual agents, it is called an additive effect; when it is greater than this sum, it is known as a synergistic effect; when one agent causes a decrease in the effect produced by another, the result is termed antagonism; and when an agent that by itself produces no effect leads to an

µg/dl (defined as the blood level used “to identify children with blood lead levels that are higher than most children’s levels,” CDC, *Blood Lead Levels in Children*, <https://perma.cc/XY53-3A7K>).

120. See, e.g., *McManaway v. KBR, Inc.*, No. H-10-1044, 2012 WL 13059744 (S.D. Tex. Aug. 12, 2012) (expert conducted a differential diagnosis for each plaintiff, in which he ruled out other causes for their injuries, and also conducted blood testing to assist with differential diagnoses regarding sodium dichromate exposure); *In re E.I. du Pont de Nemours & Co. C-8 Personal Inj. Litig.*, 348 F. Supp. 3d 680, 696 (S.D. Ohio 2016) (differential diagnosis is an “appropriate method for making a determination of causation for an individual instance of disease”) (citing *Hardyman v. Norfolk & W. Ry. Co.*, 243 F.3d 255, 260 (6th Cir. 2001)); see also *Milward v. Rust-Oleum Corp.*, 820 F.3d 469, 476 (1st Cir. 2016) (affirming exclusion of expert testimony as unreliable for failure to rule out an idiopathic diagnosis); *Kovach v. Wheeling & Lake Erie Ry. Co.*, 556 F. Supp. 3d 762, 771 (N.D. Ohio 2021) (quoting *Best v. Lowe’s Home Ctrs., Inc.*, 563 F.3d 171, 181 (6th Cir. 2009)) (noting that “doctors need not rule out every conceivable cause for their differential-diagnosis-based opinions to be admissible[.]” but they must at least consider and rule out alternative causes).

121. See generally Angelo Moretto et al., *A Framework for Cumulative Risk Assessment in the 21st Century*, 47 *Critical Revs. Toxicology* 85 (2017), <https://doi.org/10.1080/10408444.2016.1211618>; Devon C. Payne-Sturges et al., *Methods for Evaluating the Combined Effects of Chemical and Nonchemical Exposures for Cumulative Environmental Health Risk Assessment*, 15 *Int’l J. Env’t Rsch. & Pub. Health* 2797 (2018), <https://doi.org/10.3390/ijerph15122797>.

enhancement of the effect of another agent, the response is termed potentiation.¹²² These interactions can often be explained if the metabolic pathways of the agent or the toxicological mechanism by which the individual agents cause their effects are known.

Three types of toxicological approaches are pertinent to understanding the effects of mixtures of agents. One is based on the standard toxicological evaluation of common commercial mixtures, such as gasoline. The second approach is from studies in which the known toxicological effect of one agent is used to explore the mechanism of action of another agent, such as using a known specific inhibitor of a metabolic pathway to determine whether the toxicity of a second agent depends on this pathway. The third approach is based on an understanding of the basic mechanism of action of the individual components of the mixture, thereby allowing prediction of the combined effect, which can then be tested in an animal model.¹²³

The toxicologist may also need to consider whether the mixture contains agents whose effects would allow estimation of exposure to a putative causative agent that was also part of the mixture. For example, the upper boundary for exposure to benzene present at a level of 0.1% in commercial grade toluene can be based on the level of toluene that causes central nervous system effects, including coma. A worker's daily exposure of 1 part per million (ppm) benzene to this mixture would have imposed a concomitant exposure of approximately 1,000 ppm toluene—incompatible with the National Institute for Occupational Safety and Health (NIOSH) rating of Immediately Dangerous to Life and Health level for toluene of 500 ppm for fifteen minutes. Conversely, in a situation in which the worker did become woozy or comatose following exposure to a toluene-containing mixture, the toxicologist could use information about the level of toluene that caused such symptoms as a means to estimate the extent to which there was exposure to benzene or another contaminant of concern.

122. Courts have been called on to consider the issue of synergy. See *Russell v. U.S. Dep't of Labor*, No. 3:15-CV-320-DCP, 2018 WL 2054566, at *8–9 (E.D. Tenn. May 2, 2018) (finding that the Department of Labor properly considered synergistic effects of four substances in arriving at its conclusion that plaintiff's chronic obstructive pulmonary disease was not caused by exposure to the substances "individually or in combination"); but see *In re E.I. Du Pont De Nemours & Co. C-8 Personal Inj. Litig.*, 337 F. Supp. 3d 728, 747 (S.D. Ohio 2015) (noting that the expert's opinion on synergistic action between C-8 and obesity was "mere speculation" and, in this context, "synergism as a theory is unreliable and inadmissible under *Daubert* and Rule 702").

123. See generally Moretto et al., *supra* note 121; Payne-Sturges et al., *supra* note 121. The EPA has been addressing the issue of multiple exposures to different agents within a community under the heading of cumulative risk assessment. This approach is particularly of importance in dealing with environmental justice concerns. See Michael A. Callahan & Ken Sexton, *If Cumulative Risk Assessment Is the Answer, What Is the Question?*, 115 *Env't Health Persps.* 799 (2007), <https://doi.org/10.1289/ehp.9330>.

Do Humans Differ in the Extent of Susceptibility to the Particular Compound in Question? Are These Differences Relevant in This Case?

Individuals who exercise inhale more than sedentary individuals and therefore are exposed to higher doses of airborne environmental toxicants. Similarly, differences in metabolism, which are inherited or caused by external factors, such as the levels of carbohydrates in a person's diet, may result in differences in the delivery of a toxic product to the target organ.¹²⁴

Moreover, for any given level of a toxic agent that reaches a target organ, damage may be greater because of a greater response of that organ. In addition, for any given level of target-organ damage, there may be a greater impact on particular individuals. For example, an elderly individual or someone with pre-existing lung disease is less likely to tolerate a small decline in lung function caused by an air pollutant than is a healthy individual with normal lung function.

Advances in human genetics research and interactions between genetics and environmental factors are providing information about susceptibility that may be relevant to determining the likelihood that a given exposure has a specific effect on an individual.¹²⁵

Nongenetic factors may also alter the metabolism of a carcinogen. For example, an increased risk of certain cancers has been associated with obesity.¹²⁶ Obesity also can alter the metabolism of chemicals, for example through the increased levels of cytochrome P450 2E1 found in humans and laboratory animals with fatty livers.¹²⁷ A person's level of physical activity, age, sex, and genetic makeup, as well as exposure to therapeutic agents (such as prescription or over-the-counter drugs), may affect the metabolism of the compound and hence its toxicity, making the individual more susceptible to the toxic effects of a chemical exposure.

124. See generally Moretto et al., *supra* note 121; Payne-Sturges et al., *supra* note 121; see also Marissa B. Kosnik & David M. Reif, *Determination of Chemical-Disease Risk Values to Prioritize Connections Between Environmental Factors, Genetic Variants, and Human Diseases*, 379 *Toxicology & Applied Pharmacology* 114674 (2019), <https://doi.org/10.1016/j.taap.2019.114674>.

125. See, e.g., Supratim Choudhuri et al., *From Classical Toxicology to Tox21: Some Critical Conceptual and Technological Advances in the Molecular Understanding of the Toxic Response Beginning from the Last Quarter of the 20th Century*, 161 *Toxicological Scis.* 5 (2018), <https://doi.org/10.1093/toxsci/kfx186>.

126. See, e.g., Sarkees v. E.I. DuPont de Nemours & Co., 15 F.4th 584, 592–93 (2d Cir. 2021) (expert identified numerous factors that can elevate risk of bladder cancer and then sufficiently ruled them out, thereby establishing differential etiology); see also Cooper v. Smith & Nephew, Inc., 259 F.3d 194, 202 (4th Cir. 2001).

127. See, e.g., Mohamed A. Abdelmegeed et al., *Role of CYP2E1 in Mitochondrial Dysfunction and Hepatic Injury by Alcohol and Non-Alcoholic Substances*, 10 *Current Molecular Pharmacology* 207 (2017), <https://doi.org/10.2174/1874467208666150817111114>.

Has the Expert Considered Data That Contradict Their Opinion?

Multiple avenues of deductive reasoning based on scientific data lead to acceptance of causation in any field, particularly in toxicology. However, the basis for this deductive reasoning is also one of the most difficult aspects of causation to describe quantitatively. If animal studies, pharmacological research on mechanisms of toxicity, in vitro tissue studies, and epidemiological research all document toxic effects of exposure to a compound, an expert's opinion about causation in a particular case is much more likely to be true.¹²⁸

The more difficult problem is how to evaluate conflicting research results. When different research studies reach different conclusions regarding toxicity, the expert must be asked to explain how those results have been taken into account in the formulation of the expert's opinion.

Expert Qualifications

The basis of the toxicologist's expert opinion in a specific case is a thorough review of the research literature and treatises concerning effects of exposure to the chemical at issue. To arrive at an opinion, the expert assesses the strengths and weaknesses of the research studies. The expert also bases an opinion on fundamental concepts of toxicology relevant to understanding the actions of chemicals in biological systems.

As the following series of questions indicates, no single academic degree, research specialty, or career path qualifies an individual as an expert in

128. See, e.g., *Waite v. All Acquisition Corp.*, 194 F. Supp. 3d 1298, 1313, 1315 (S.D. Fla., 2016) (holding that the weight-of-evidence approach used by the expert was “thoroughly reasoned and based on sound methodology” and explaining that the approach “requires consideration of all available scientific evidence, including epidemiology, toxicology (animal studies), cellular studies (in vitro), and molecular biology”); *In re Flint Water Cases*, No. 17-10164, 2021 WL 5631706, at *7–8 (E.D. Mich. Dec. 1, 2021) (finding expert's opinion on general causation reliable where the opinion was based on “reasonable scientific inferences” from “the great weight of the evidence” and where the expert “cite[d] to peer-reviewed, epidemiological studies which account for confounding variables”); *In re Abilify (Aripiprazole) Prods. Liab. Litig.*, 299 F. Supp. 3d 1291, 1311 (N.D. Fla. 2018) (explaining that the “‘weight of the evidence’ approach to analyzing causation can be considered reliable, provided the expert considers all available evidence carefully and explains how the relative weight of the various pieces of evidence led to his conclusion . . . , it is crucial that the expert describe each step in the process by which he gathered and assessed the scientific evidence”). But see *Daniels-Feasel v. Forest Pharms., Inc.*, No. 17 CV 4188-LTS-JLC, 2021 WL 4037820, at *5 (S.D.N.Y. Sept. 3, 2021), *aff'd*, No. 22-146, 2023 WL 4837521 (2d Cir. July 28, 2023) (quoting *In re Abilify*, 299 F. Supp. 3d at 1311) (expert's opinion that relies on cherry-picked data is not reliable). See generally Liesa L. Richter & Daniel J. Capra, *The Admissibility of Expert Testimony*, in this manual.

toxicology. Toxicology is a heterogeneous field. A number of indicia of expertise can be explored, however, that are relevant to both the admissibility and weight of the proffered expert opinion.

Does the Proposed Expert Have an Advanced Degree in Toxicology, Pharmacology, or a Related Field? If the Expert Is a Physician, Are They Board Certified in Toxicology or in a Field Such as Occupational Medicine?

A graduate degree in toxicology demonstrates that the proposed expert has a substantial background in the basic issues and tenets of toxicology. Many universities have established graduate programs in toxicology. These programs are administered by the faculties of medicine, pharmacology, pharmacy, or public health.

In addition to toxicologists with a specific Ph.D. in toxicology, many highly qualified toxicologists are physicians or hold doctoral degrees in related disciplines (e.g., veterinary medicine, pharmacology, biochemistry, environmental health, or industrial hygiene). For a person with this type of background, a single course in toxicology is unlikely to provide sufficient background for developing expertise in the field. Additional specific training and board certification in toxicology is highly relevant to evaluating the credentials of a toxicologist.

A proposed expert should be able to demonstrate an understanding of the discipline of toxicology, including statistics, toxicological research methods, and disease processes. A physician without particular training or experience in toxicology is unlikely to have sufficient background to evaluate the strengths and weaknesses of toxicological research. Most practicing physicians have little knowledge of environmental and occupational medicine.¹²⁹ Subspecialty physicians may have particular knowledge of a cause-and-effect relationship (e.g., pulmonary physicians have knowledge of the relationship between asbestos exposure and asbestosis); anesthesiology is largely focused on the effects of certain chemicals; and certified addiction medicine physicians focus on chemical agents that are not pharmaceuticals or are illegally used pharmaceuticals, such as

129. For recent documentation of how rarely an occupational history is obtained, see Barry J. Politi et al., *Occupational Medical History Taking: How Are Today's Physicians Doing? A Cross-Sectional Investigation of the Frequency of Occupational History Taking by Physicians in a Major U.S. Teaching Center*, 46 J. Occupational Env't Med. 550 (2004), <https://doi.org/10.1097/01.jom.0000128153.79025.e4>.

fentanyl.¹³⁰ Physicians who receive additional training in classic toxicology include those who are certified by the American Board of Preventive Medicine in medical toxicology, addiction medicine, or occupational and environmental medicine.¹³¹ Knowledge of toxicology is particularly strong among occupational health specialists who work in the chemical, petrochemical, and pharmaceutical industries, in which the surveillance of workers exposed to chemicals is a major responsibility.¹³²

130. Courts look at specific circumstances in determining whether a treating physician may testify as a fact witness or an expert witness. *See, e.g.,* *Mueller v. Chugach Fed. Solutions, Inc.*, No. 12-S-00624-NE, 2014 WL 2891030, at *3–4 (N.D. Ala. June 25, 2014) (permitting treating physician to testify as fact witness and offer his “opinion regarding the cause of death . . . based upon his personal experience of treating” the decedent); *but see* N.K. by Bruestle-Kumra v. Abbott Lab’s, 731 F. App’x 24, 26 (2d Cir. 2018) (noting that a treating physician may not testify as a fact witness and must “be admitted as a Rule 702 expert witness to provide expert testimony on specific causation”). Some courts have allowed treating physicians to testify regarding causation because it could have been formed within the scope of treatment); *Burton v. Am. Cyanamid*, 362 F. Supp. 3d 588, 599 (E.D. Wis. 2019) (allowing treating physician to testify regarding lead exposure and causation because “his professional training and his experience treating . . . patients with elevated levels are sufficient to qualify him to give testimony”). *Contra* *Zellers v. NexTech Ne., LLC*, 533 F. App’x 192, 198 (4th Cir. 2013) (excluding testimony of treating physician as to causation because, *inter alia*, the physician was a neurologist who did not have “specialized training in the field of toxicology”); *Higgins v. Koch Dev. Corp.*, 794 F.3d 697, 704–05 (7th Cir. 2015) (excluding treating physician opinion regarding causation because the treating physician had no “training in toxicology”).

Treating physicians also become involved in considering cause-and-effect relationships when they are asked whether a patient can return to a situation in which an exposure has occurred. The answer is obvious if the cause-and-effect relationship is clearly known. However, this relationship is often uncertain, and the physician must consider the appropriate advice. In such situations, the physician will tend to give advice as though the causality was established, both because it is appropriate caution and because of fears concerning medicolegal issues.

131. Before 1990, the American Board of Medical Toxicology certified physicians. But beginning in 1990, medical toxicology became a subspecialty board under the American Board of Emergency Medicine, the American Board of Pediatrics, and the American Board of Preventive Medicine, as recognized by the American Board of Medical Specialties.

132. Clinical ecologists, another group of physicians, have offered opinions regarding multiple chemical hypersensitivity and immune system responses to chemical exposures. These physicians generally have a background in the field of allergy, not toxicology, and their theoretical approach is derived in part from classic concepts of allergic responses and immunology. This theoretical approach has often led clinical ecologists to find cause-and-effect relationships or low-dose effects that are not generally accepted by toxicologists. Clinical ecologists often belong to the American Academy of Environmental Medicine.

In 1991, the Council on Scientific Affairs of the American Medical Association concluded that until “accurate, reproducible, and well-controlled studies are available . . . multiple chemical sensitivity should not be considered a recognized clinical syndrome.” Council on Sci. Affairs, Am. Med. Ass’n, Council Report on Clinical Ecology 6 (1991). Multiple chemical sensitivity has been held by courts to be “a controversial diagnosis that has been excluded under *Daubert* as unsupported

Has the Proposed Expert Been Certified by the American Board of Toxicology, or Do They Belong to a Professional Organization, Such as the Academy of Toxicological Sciences or the Society of Toxicology?

As of August 2022, more than 2,500 individuals had received board certification from the American Board of Toxicology, Inc., and thus can use the title of “Diplomate of the American Board of Toxicology” (DABT). To sit for the examination, the candidate must be involved full time in the practice of toxicology, including designing and managing toxicological experiments or interpreting results and translating them to identify and solve human and animal health problems. Diplomates must be recertified every five years. The Academy of Toxicological Sciences (ATS) was formed to provide credentials in toxicology through peer review only. It does not administer examinations for certification. As of 2022, approximately 290 individuals had been elected as Fellows of ATS.

The Society of Toxicology (SOT), the major professional organization for the field of toxicology, was founded in 1961 and has grown dramatically in recent years. In 2022 the SOT had approximately 800 members.¹³³ Criteria for membership is based either on peer-reviewed publications or on the active practice of toxicology. Physician toxicologists can join the American College of Medical Toxicology and the American Academy of Clinical Toxicology as well as the SOT. There are also societies of forensic toxicology, such as the International Association of Forensic Toxicologists. Other organizations in the field are the American College of Toxicology, for which experience in the active practice of toxicology is the major membership criterion; the International Society of Regulatory Toxicology and Pharmacology; and the Society for Occupational and Environmental Health. For membership, the last two organizations require only the payment of dues.

by sound scientific reasoning or methodology.” *Madej v. Maiden*, 951 F.3d 364, 374–75 (6th Cir. 2020) (quoting *Summers v. Mo. Pac. R.R. Sys.*, 132 F.3d 599, 603 (10th Cir. 1997)) (collecting cases). The *Madej* court recognized that the “mountain of precedent” regarding multiple chemical sensitivity consists mostly of cases “over a decade old[.]” but reiterated that it has not been “shown that more recent scientific advancements have led the scientific community to come to accept a multiple-chemical-sensitivity diagnosis.” 951 F.3d at 375. The “diagnosis remains unrecognized by the American Medical Association and unlisted in the World Health Organization’s International Classification of Diseases.” *Id.*

133. As of 2024, there are twenty-nine specialty sections of the SOT that represent the different specialty areas involved in understanding the wide range of toxic effects associated with exposure to chemical and physical agents. These sections include Mechanisms, Molecular and Systems Biology, Inhalation and Respiratory, Metals, Neurotoxicology, Carcinogenesis, Risk Assessment, Biological Modeling, Immunotoxicology, and Sustainable Chemicals through Contemporary Toxicology, to name a few.

What Other Criteria Does the Proposed Expert Meet?

Toxicology, like many other disciplines that transcend both basic and applied science, has in recent decades gone through professionalization. This includes development of organizations that have different levels of expertise and recognition in the field. Understanding these organizations, including their criteria for membership, is pertinent to evaluating expertise in the field.

The success of academic scientists in toxicology, as in other biomedical sciences, usually is measured by the following types of criteria: the quality and number of peer-reviewed publications, the ability to compete for research grants, service on scientific advisory panels, and university appointments.

Publication of articles in peer-reviewed journals indicates an expertise in toxicology. The number of articles, their topics, and whether the individual is the principal or senior author are important factors in determining the expertise of a toxicologist.¹³⁴

Most research grants from government agencies and private foundations are highly competitive. Successful competition for funding and publication of the research findings indicate competence in an area.

Selection for local, national, and international regulatory advisory panels usually implies recognition in the field. Examples of such panels are the NIH Study Sections considering toxicology-related grant proposals, and panels convened by the EPA, FDA, WHO, and IARC. Recognized industrial organizations, including the American Petroleum Institute and the Electric Power Research Institute; public interest groups, such as the Environmental Defense Fund and the Natural Resources Defense Council; and public-private nonprofit partnerships, such as the Health and Environmental Sciences Institute (HESI) and the Health Effects Institute (HEI), employ toxicologists directly and as consultants, and enlist academic toxicologists to serve on advisory panels. Because of a growing interest in environmental issues, the demand for scientific advice has outgrown the supply of available toxicologists. It is thus common for reputable toxicologists to serve on advisory panels.

Finally, a tenured or tenure-track university appointment in toxicology, risk assessment, or a related field signifies substantial expertise, particularly if the

134. Examples of reputable, peer-reviewed journals are the *Journal of Toxicology and Environmental Health*; *Toxicological Sciences*; *Toxicology and Applied Pharmacology*; *Science*; *British Journal of Industrial Medicine*; *Clinical Toxicology*; *Archives of Environmental Health*; *Journal of Occupational and Environmental Medicine*; *Annual Review of Pharmacology and Toxicology*; *Teratogenesis, Carcinogenesis and Mutagenesis*; *Fundamental and Applied Toxicology*; *Inhalation Toxicology*; *Biochemical Pharmacology*; *Toxicology Letters*; *Environmental Research*; *Environmental Health Perspectives*; *International Journal of Toxicology*; *Human and Experimental Toxicology*; *Regulatory Toxicology and Pharmacology*; and *American Journal of Industrial Medicine*.

university has a graduate education program in that area. However, some universities offer adjunct or affiliate faculty positions to local professionals (consultants, government employees, etc.), often to assist in teaching entry-level toxicology courses. Although many well-qualified toxicologists working in the private sector or government agencies often have adjunct academic appointments, such an appointment does not necessarily confer the same level of expertise or toxicology credentials as a regular faculty appointment.

Acknowledgments

The authors gratefully appreciate the excellent research assistance provided by Stephanie N. Patton, associate, Buchanan Ingersoll & Rooney, P.C.; Lyndsey Arneson Field, Wake Forest Law School class of 2024; and Marvin Astrada, senior research associate, Federal Judicial Center, in the preparation of this fourth edition of the reference guide.

Glossary of Terms

absorption. The biological process of taking up of a chemical into the body orally, through inhalation, or through skin exposure.

acute toxicity. An immediate toxic response following a single or short-term exposure to a toxic substance.

additive effect. When exposure to more than one toxic agent results in the same effect as would be predicted by the sum of the effects of exposure to the individual agents.

antagonism. When exposure to one toxic agent causes a decrease in the effect produced by another toxic agent.

benchmark dose (BMD). The benchmark dose is determined on the basis of dose–response modeling and is defined as the exposure associated with a specified low incidence of risk, generally in the range of 1% to 10%, of a health effect, or the dose associated with a specified measure or change of a biological effect. Frequently, the statistical lower 10% confidence limit on the estimate of the BMD is used for regulatory standard setting, and is referred to as the Benchmark Dose Low 10%, or BMDL₁₀.

bioassay. A test for measuring the toxicity of an agent by exposing laboratory animals to the agent and observing the effects.

biological monitoring. Measurement of toxic agents or the results of their metabolism in biological materials, such as blood, urine, expired air, or biopsied tissue, to test for exposure to the toxic agents, or the detection of physiological changes that are due to exposure to toxic agents.

biologically plausible theory. A biological explanation for the relationship between exposure to an agent and adverse health outcomes. This is usually associated with an understanding of how the chemical causes its toxic effect (see note 49 above and accompanying text).

carcinogen. A chemical substance or other agent that causes cancer.

carcinogenicity bioassay. Limited or long-term tests using laboratory animals to evaluate the potential carcinogenicity of an agent.

chronic toxicity. A toxic response to long-term exposure or dosing (typically more than six months and up to two years for laboratory animal studies) with an agent.

clinical ecologists. Physicians who believe that exposure to certain chemical agents can result in damage to the immune system, causing multiple-chemical hypersensitivity and a variety of other disorders. Clinical ecologists often have a background in the field of allergy, not toxicology, and their theoretical approach is derived in part from classic concepts of allergic

responses and immunology. There has been much resistance in the medical community to accepting their claims.

clinical toxicology. The study and treatment of humans exposed to chemicals and the quantification of resulting adverse health effects. Clinical toxicology includes the application of pharmacological principles to the treatment of chemically exposed individuals and research on measures to enhance elimination of toxic agents.

compound. In chemistry, the combination of two or more different elements in definite proportions, which when chemically combined acquire properties different from those of the original elements.

confounding factors. Variables that are related to both exposure to a toxic agent and the outcome of the exposure. A confounding factor can obscure the relationship between the toxic agent and the adverse health outcome associated with that agent.

developmental toxicology. The processes by which a chemical agent causes toxic effects to the developing embryo and fetus and also during early life stages after birth when the brain and endocrine system continue to develop.

differential diagnosis. A physician's consideration of alternative diagnoses that may explain a patient's condition.

direct-acting agents. Agents that cause toxic effects without metabolic activation or conversion.

distribution. Movement of a toxic agent throughout the organ systems of the body (e.g., the liver, kidney, bone, fat, and central nervous system). The rate of distribution is usually determined by the blood flow through the organ and the ability of the chemical to pass through the cell membranes of the various tissues.

dose, dosage. A product of both the concentration of a chemical or physical agent and the duration and frequency of exposure.

dose–response curve. A graphic representation of the relationship between the dose of a chemical administered and the effect(s) produced. Each specific endpoint of toxicity may have its own dose–response curve.

dose–response relationships. The extent to which a living organism responds to specific doses of a toxic substance. The more time spent in contact with a toxic substance, or the higher the dose, the greater the organism's response. For example, a small dose of carbon monoxide will cause drowsiness; a large dose can be fatal.

epidemiology. The study of the occurrence and distribution of disease among people. Epidemiologists study groups of people to discover the cause of a disease, or where, when, and why disease occurs.

epigenetic. The study of how cells control gene activity without changing the DNA sequence. Epigenetics research has identified heritable changes in the functions of DNA that are not associated with a change in the DNA sequence. Chemicals found in the diet, workplace, and environment have been found to cause epigenetic changes that may contribute to adverse effects, including cancers.

etiology. A branch of medical science concerned with the causation of diseases.

excretion. The process by which toxicants are eliminated from the body, including through the kidneys and urinary tract, the liver and biliary system, fecal excretion, exhalation of the agent via the lungs, and secretion into sweat, saliva, and breast milk (lactation, which can result in an important route of exposure for nursing infants).

exposure. The intake into the body of a hazardous material. The main routes of exposure to substances are through the skin, mouth, and lungs. Exposure via injection (intravenous, intramuscular, subcutaneous) can occur for some drugs, especially drugs of abuse.

exposure assessment. The second step in the process of chemical risk assessment, in which quantitative estimates of the potential dose and duration of exposure to a specific chemical or chemical mixture will occur during a specified set of circumstances. It includes the consideration of concentration of a chemical in specific media (air, food, water, skin contact), amount of contact (food consumption, water consumption, inhalation rate), and frequency and duration of exposure for each route.

extrapolation. The process of estimating unknown values from known values.

genome, genomics. The genetic content of a cell, composed of DNA and related proteins that make up the twenty-three pairs of chromosomes in every living human cell. The human genome is composed of approximately 24,000 genes, each consisting of thousands of individual nucleotides (the building blocks of DNA). The sequence of the DNA in a gene dictates its function. The DNA of a gene is used by a cell to make proteins, which are the functional units of a cell. Changes in DNA sequence (a mutation) can result in irreversible changes in the function of a cell, leading to a wide variety of pathology, including cancers and birth defects.

good laboratory practice (GLP). Codes developed by the federal government in consultation with the laboratory testing industry that govern many aspects of laboratory standards.

hazard. The inherent ability of a chemical to cause (a) specific type(s) of toxic effect on biological organisms. Hazard differs from risk in that there is no quantitative consideration of dose.

hazard assessment. The first step in the process of risk assessment, which evaluates the potential of a specific chemical, or mixture of chemicals, to cause harm.

hazard identification. In risk assessment, the qualitative analysis of all available experimental animal and human data to determine whether and at what dose an agent is likely to cause toxic effects.

hydrologists, hydrogeologists. Scientists who specialize in the movement of ground and surface waters and the distribution and movement of contaminants in those waters.

immunotoxicology. A branch of toxicology concerned with the effects of toxic agents on the immune system.

indirect-acting agents. Agents that require metabolic activation or conversion before they produce toxic effects in living organisms.

inhalation toxicology. The study of the effect of toxic agents that are absorbed into the body through inhalation, including their effects on the respiratory system.

in silico methodology. Computational models that investigate toxicological and pharmacological hypotheses using tools such as data mining, homology models, machine learning, pharmacophores, quantitative structure–activity relationships, and network analysis tools.

in vitro. A research or testing methodology that uses living cells in an artificial or test tube system, or that is otherwise performed outside of a living organism.

in vivo. A research or testing methodology that uses living organisms.

lethal dose 50 (LD₅₀). The dose at which 50% of laboratory animals die within days to weeks, generally after the administration of a single dose.

lifetime bioassay. A bioassay in which doses of an agent are given to experimental animals throughout their lifetime. (See bioassay.) For carcinogenicity and most chronic toxicity studies, the standardized period for a lifetime bioassay in rats and mice is two years.

linearized non-threshold (LNT) model. A mathematical modeling approach that is used to extrapolate dose–response data to doses below the observed range, with assumptions that (1) the cancer risk is directly proportional to the dose at all doses, and thus (2) there is no threshold (a dose below which the response is zero).

lowest observed adverse effect level (LOAEL). The lowest dose or concentration of a chemical that causes observable adverse effects in experimental toxicology studies that use multiple doses.

maximum tolerated dose (MTD). The highest dose of an agent to which an organism can be exposed without it causing death or significant overt toxicity.

mechanism (or mode) of action. The specific molecular, biochemical, and cellular processes that are altered by the toxic substance, giving rise to the observed adverse effect(s). (See also biologically plausible theory.)

metabolism. The sum total of the biochemical reactions that a chemical produces in an organism.

metabolome. The total number of small molecules, both endogenous and exogenous, that can be measured in the blood. Metabolomics is useful to identify the presence of multiple potentially toxic substances in the blood, including the metabolites of the toxic substances.

molecular toxicology. The study of how toxic agents interact with cellular molecules, including DNA, proteins, and lipids (fats).

multiple-chemical hypersensitivity. A physical condition whereby individuals react to many different chemicals at extremely low exposure levels.

multistage model. A biochemical and statistical model for understanding certain diseases, including some cancers, based on the postulate that more than one event is necessary for the onset of disease.

mutagen. A substance that causes physical changes in chromosomes or biochemical changes in genes.

mutagenesis. The process by which agents cause changes in chromosomes and genes.

neurotoxicology. A branch of toxicology concerned with the effects of exposure to toxic agents on the central nervous system.

no observed adverse effect level (NOAEL). The highest level of exposure to an agent at which no effect is observed. It is the experimental equivalent of a threshold. See threshold.

one-hit hypothesis. A hypothesis of cancer risk in which each molecule of a chemical mutagen has a possibility, no matter how tiny, of mutating a gene in a manner that may lead to tumor formation or cancer. See also linearized non-threshold model.

PFAS. A widely used abbreviation to represent a class of chemicals called per- and polyfluoroalkyl substances. Such compounds provide a nonstick and stain-resistant coating to a wide variety of consumer products, such as nonstick cookware, carpeting, apparel, upholstery, personal care products, and cosmetics, and were also used in firefighting foams.

pharmacokinetics. A mathematical approach that expresses the movement of a toxic agent through the organ systems of the body, including the distribution to the target organ and to its ultimate fate via excretion pathways. It measures rate constants for each step in the process.

point of departure (POD). A term used to identify a point on the lower portion of a dose–response curve from which extrapolations are used to identify a hypothetical response level at lower doses. PODs are usually either a NOAEL, LOAEL, or BMD. These values, which are experimentally derived within the range of observations of a specified effect (e.g., birth defect, liver injury, tumor occurrence), are then divided by specified safety or uncertainty factors to arrive at a Reference Dose (RfD) or Acceptable Daily Intake (ADI).

potentiation. The process by which the addition of one agent, which by itself has no toxic effect, increases the toxicity of another agent when exposure to both agents occurs simultaneously.

reference dose (RfD). A term used by some regulatory agencies, such as the EPA, to identify a daily dose that is expected to be without harm. It is conceptually identical to the phrase acceptable daily intake (ADI), which is the amount of a chemical to which a person can be exposed on a daily basis over an extended period of time (usually a lifetime) without suffering a deleterious effect.

reproductive toxicology. The study of the effect of toxic agents on male and female reproductive systems, including sperm, ova, and offspring.

risk assessment. The use of scientific evidence to estimate the likelihood of adverse effects on the health of individuals or populations from exposure to hazardous materials and conditions. Chemical risk assessment includes: (1) hazard assessment, (2) exposure assessment, (3) dose–response modeling, and (4) risk characterization.

risk characterization. The final step in the risk assessment process, which summarizes information about an agent, including the strengths and limitations of the available toxicity data, and evaluates it in order to estimate the risks it poses.

safety assessment. Toxicological research that tests the toxic potential of a chemical in vivo or in vitro using standardized techniques required by governmental regulatory agencies or other organizations. The term is similar to risk assessment of chemical hazards, but is most often used in association with pharmaceuticals, food additives, and other chemicals for which public benefits are expected. It considers both the risks and the benefits of the chemical in deriving an acceptable level of risk (i.e., a “safe” dose).

safety factor (SF). See uncertainty factor.

structure–activity relationships (SAR). An in silico method used by toxicologists to predict the toxicity of new chemicals by comparing their chemical structures with those of compounds with known toxic effects.

synergistic effect. An effect by which two toxic agents acting together have an effect greater than that predicted by adding together their individual effects.

target organ. The organ system that is affected by a particular toxic agent.

target-organ dose. The dose to the organ that is affected by a particular toxic agent.

teratogen. An agent that changes eggs, sperm, or embryos, thereby increasing the risk of birth defects.

teratogenic. The ability to produce birth defects. Teratogenic effects do not pass to future generations. See teratogen.

threshold. The dose level above which effects will occur and below which no effects are expected to occur. See no observed adverse effect level.

toxic. Of, relating to, or caused by a poison—or a poison itself.

toxic agent or toxicant. An agent or substance that causes disease or injury.

toxicokinetics. Toxicokinetics is the description of the rate at which a chemical will enter an organism, and what occurs in terms of absorption, distribution, metabolism, and excretion (ADME). The term is used principally in describing ADME characteristics for chemicals that are not used as pharmaceuticals. The term pharmacokinetics (see above) describes this same process for chemicals used for therapeutic benefit (drugs).

toxicology. The science of the nature and effects of poisons, their detection, and the treatment of their adverse effects.

toxins. Toxic substances that are produced by biological organisms (plants, animals, fungi, bacteria, etc).

uncertainty factor (UF). These are values, usually a factor of ten, which are used to provide ample public health protection in determining “safe doses” (e.g., ADIs, RfDs) when extrapolating dose information obtained in experimental animals to human populations.

xenobiotic. A chemical that is foreign to the body. Sometimes used synonymously with “toxicant” or “toxic chemical.”

General References on Toxicology and Chemical Risk Assessment

- Casarett & Doull's Toxicology: The Basic Science of Poisons (Curtis D. Klaassen ed., 9th ed. 2019).
- Environmental Health: From Global to Local (Howard Frumkin ed., 3d ed. 2016).
- Encyclopedia of Toxicology (Phillip Wexler ed., 3d ed. 2014).
- Hayes' Principles and Methods of Toxicology (A. Wallace Hayes & Claire L. Kruger eds., 6th ed. 2014).
- National Research Council, Applications of Toxicogenomic Technologies to Predictive Toxicology and Risk Assessment (2007).
- National Research Council, New Approach Methods (NAMs) for Human Health Risk Assessment: Proceedings of a Workshop—in Brief (2022).
- National Research Council, Science and Decisions: Advancing Risk Assessment (2009).
- National Research Council, Toxicity Testing in the 21st Century: A Vision and a Strategy (2007).
- National Research Council, Using 21st Century Science to Improve Risk-Related Evaluations (2017).
- National Research Council, Understanding Pathways to a Paradigm Shift in Toxicity Testing and Decision-Making: Proceedings of a Workshop (2018).
- Principles of Toxicology*, in Comprehensive Toxicology (vol. 1, D. L. Eaton vol. ed., Charlene McQueen series ed., 3d ed. 2017).

Reference Guide on Medical Testimony

JOHN B. WONG, LAWRENCE O. GOSTIN, AND
OSCAR A. CABRERA

John B. Wong, M.D., is Vice Chair of Academic Affairs and Chief of the Division of Clinical Decision Making in the Department of Medicine at Tufts Medical Center and Professor of Medicine at Tufts University School of Medicine.

Lawrence O. Gostin, J.D., is Linda D. and Timothy J. O'Neill Professor of Global Health Law and Faculty Director of the O'Neill Institute for National and Global Health Law at Georgetown University Law Center.

Oscar A. Cabrera, Abogado, LL.M., is Deputy Director of the O'Neill Institute for National and Global Health Law and an Adjunct Professor of Law at Georgetown University Law Center.

CONTENTS

Introduction, 1108

Medical Testimony Introduction, 1109

 Medical Versus Legal Terminology, 1109

 Testimony by Physicians, 1113

 Medical Expertise and Admissibility, 1115

Medical Care, 1115

 Medical Education and Training, 1115

 Medical School, 1115

 Postgraduate Training, 1117

 Licensure and Credentialing, 1117

 Continuing Medical Education, 1118

 Organization of Medical Care, 1119

 Patient Care, 1121

 Goals, 1121

 Diagnostic Process, 1122

Medical Decision-Making, 1124

 Clinical-Reasoning Process, 1125

 Hypothetico-Deductive Reasoning, 1125

 Heuristics: Intuitive Pattern Recognition, 1127

 Diagnostic Reasoning, 1128

 Symptoms and Signs, 1129

 Bayes' Rule, 1129

 Probabilistic Reasoning, 1130

Prosecutor’s Fallacy, 1133	
Multiple Diseases and Multiple Tests, 1135	
Genomic Analyses, 1135	
Positivity Criterion, 1136	
Bayesian vs. Frequentist Statistics, 1136	
Causal Reasoning, 1138	
Causal Inference Methods, 1140	
Treatment Decision-Making, 1142	
Treatment Thresholds, 1142	
Testing Decision-Making, 1143	
Testing Thresholds, 1143	
Screening, 1144	
Diagnostic Testing, 1146	
Biomarker or Image Testing, 1147	
Variation and Standards in Medicine, 1147	
Variation in Medical Care, 1147	
Evidence-Based Medicine, 1149	
Hierarchy of Medical Evidence, 1150	
Guidelines, 1152	
Uncertainty and Tradeoffs in Therapeutic Decision-Making Evidence, 1155	
Biases, 1161	
Cognitive psychology, 1161	
Race and ethnicity, 1162	
Estimating kidney function with race and without, 1163	
Algorithms as standards, 1164	
Algorithmic fairness and clinical prediction, 1165	
Informed Consent, 1167	
Principles, 1167	
Standards, 1168	
Patient care, 1169	
Patient preferences, 1169	
Randomized trials, 1170	
Health information sources, 1170	
Risk Communication, 1171	
Types of risk-communication estimates, 1172	
Breast-cancer screening communication, 1173	
Technologies, 1174	
Data collection and use, 1175	
Shared Decision-Making, 1176	
Summary and Future Directions, 1177	
Glossary of Terms, 1179	
References on Medical Testimony, 1183	

Reference Guide on Medical Testimony

FIGURES

1. The diagnostic process, 1123
2. Medical deductive and inductive inference, 1126
3. Clinical-reasoning process from chief complaint to therapeutic decisions, 1127
4. Treatment threshold, 1142
5. Test and test–treatment thresholds, 1144
6. Mutual incompatibility of fairness criteria, 1166

TABLES

1. A Tabular Form of Bayes' Rule for Test Interpretation, 1130
2. A Tabular Form of Bayes' Rule for Diagnosis, 1131
3. Probability of Two Deaths from SIDS vs. Homicides, 1134

Introduction

Physicians are a common sight in today's courtroom. A survey of federal judges published in 2002 found that medical and mental health experts constitute more than 40% of testifying experts.¹ Medical evidence is a common element in product-liability suits,² workers' compensation disputes,³ medical malpractice suits,⁴ and personal-injury cases.⁵ Medical testimony may also be critical in certain kinds of criminal cases.⁶ This reference guide introduces types of evidence that physicians use to make judgments as treating physicians or as experts retained by one of the parties in a case for federal and state judges, emphasizing the tools and basic concepts of diagnostic reasoning and clinical decision-making and highlighting the challenges in testifying as medical experts. The "Medical Care" and "Medical Decision-Making" sections of this guide explain in detail the practice of medicine, including medical education and training, the structure and organization of healthcare, the elements of patient care, and, most importantly,

1. Joe S. Cecil, *Ten Years of Judicial Gatekeeping Under Daubert*, 95 Am. J. Pub. Health S74–S80 (2005), <https://doi.org/10.2105/AJPH.2004.044776>.

2. See, e.g., *In re Bextra & Celebrex Mktg. Sales Pracs. & Prod. Liab.*, 524 F. Supp. 2d 1166 (N.D. Cal. 2007) (thoroughly reviewing the proffered testimony of plaintiff's expert cardiologist and neurologist in a products-liability suit alleging that defendant's arthritis pain medication caused serious cardiovascular injury).

3. See, e.g., *AT&T Alascom v. Orchitt*, 161 P.3d 1232 (Alaska 2007) (affirming the decision of the state workers' compensation board and rejecting appellant's challenges to worker's experts); *Butts v. Dept. of Labor & Workforce Dev.*, 467 P.3d 231 (Alaska 2020) (reviewing the use of medical evidence to determine compensation).

4. See, e.g., *Schneider ex rel. Est. of Schneider v. Fried*, 320 F.3d 396 (3d Cir. 2003) (allowing a physician to testify in a malpractice case regarding whether administering a particular drug during angioplasty was within the standard of care); *Ellison v. United States*, 753 F. Supp. 2d 468 (E.D. Pa. 2010) (holding that oral surgeon's testimony was relevant and reliable); *Hemsley v. Langdon*, 909 N.W.2d 59 (Neb. 2018) (holding that the district court did not abuse its discretion by admitting doctor's expert testimony); *Gonzales v. Neb. Pediatric Prac., Inc.*, 955 N.W.2d 696 (Neb. 2021) (admitting the testimony of a family and emergency-room physician to prove another physician's misdiagnosis).

5. See, e.g., *Epp v. Lauby*, 715 N.W.2d 501 (Neb. 2006) (detailing the opinions of two physicians regarding whether plaintiff's fibromyalgia resulted from an automobile accident with two defendants); *Marsh v. Valyou*, 977 So. 2d 543 (Fla. 2007) (analyzing expert testimony linking a car accident to fibromyalgia); *Britt v. Wal-Mart Stores E., LP*, 599 F. Supp. 3d 1259 (S.D. Fla. 2022) (deciding that a surgeon could testify competently as to his opinions on whether or not plaintiff had sustained a permanent injury based on the information disclosed during the course of plaintiff's treatment).

6. See *State v. Price*, 171 P.3d 1223 (Mont. 2007) (discussing an assault case in which a physician testified regarding the potential for a stun gun to cause serious bodily harm); *People v. Unger*, 749 N.W.2d 272 (Mich. Ct. App. 2008) (a second-degree murder case involving testimony of a forensic pathologist and neuropathologist); *State v. Greene*, 951 So.2d 1226 (La. Ct. App. 2007) (regarding a child-sexual-battery and child-rape case involving the testimony of a board-certified pediatrician); *Unger v. Bergh*, 742 F. App'x 55 (6th Cir. 2018) (discussing the handling of expert testimony by trial counsel).

the processes of diagnostic reasoning and medical judgment that physicians use to diagnose and treat their patients. Special attention is given to the physician-patient relationship and to the types of evidence that physicians use to make medical judgments. Following the introduction, the section titled “Medical Testimony Introduction” identifies a few overarching theoretical issues that courts face in translating the methods and techniques customary in the medical profession in a manner that will serve the court’s inquiry. In an effort to make each issue more salient, examples from case law are offered when they are illustrative.

Medical Testimony Introduction

Medical Versus Legal Terminology

Because medical testimony is common in the courtroom generally and is indispensable to certain cases, courts have employed some medical terms in ways that differ from their use by the medical profession. Differential diagnosis, for example, is an accepted method that a medical expert may employ to offer expert testimony that satisfies Federal Rule of Evidence 702.⁷ In the legal context it refers to a technique “in which a physician first rules in all scientifically plausible causes of plaintiff’s injury, then rules out least plausible causes of injury until the most likely cause remains, thereby reaching a conclusion as to whether defendant’s product caused injury.”⁸ But in the medical context, differential diagnosis

7. See Liesa L. Richter & Daniel J. Capra, *The Admissibility of Expert Testimony*, in this manual.

8. *Wilson v. Taser Int’l, Inc.*, 303 F. App’x 708 (11th Cir. 2008) (“[N]onetheless, Dr. Meier did not perform a differential diagnosis or any tests on Wilson to rule out osteoporosis and these corresponding alternative mechanisms of injury. Although a medical expert need not rule out every possible alternative in order to form an opinion on causation, expert opinion testimony is properly excluded as unreliable if the doctor ‘engaged in very few standard diagnostic techniques by which doctors normally rule out alternative causes and the doctor offered no good explanation as to why his or her conclusion remained reliable’ or if ‘the defendants pointed to some likely cause of the plaintiff’s illness other than the defendants’ action and [the doctor] offered no reasonable explanation as to why he or she still believed that the defendants’ actions were a substantial factor in bringing about that illness.’”); *Williams v. Allen*, 542 F.3d 1326, 1333 (11th Cir. 2008) (“Williams also offered testimony from Dr. Eliot Gelwan, a psychiatrist specializing in psychopathology and differential diagnosis. Dr. Gelwan conducted a thorough investigation into Williams’ background, relying on a wide range of data sources. He conducted extensive interviews with Williams and with fourteen other individuals who knew Williams at various points in his life.”) (involving a capital murder defendant petitioning for habeas corpus offering a supporting expert witness); *Bland v. Verizon Wireless, L.L.C.*, 538 F.3d 893, 897 (8th Cir. 2008) (“Bland asserts Dr. Sprince conducted a differential diagnosis which supports Dr. Sprince’s causation opinion. We have held, ‘a medical opinion about causation, based upon a proper differential diagnosis, is sufficiently reliable to satisfy *Daubert*.’ A ‘differential diagnosis [is] a technique that identifies the cause of a medical condition by eliminating the likely causes until the most probable cause is isolated.’”) (citations

refers to a set of diseases that physicians consider as possible causes for the symptoms a patient is suffering or signs that a patient exhibits.⁹ By identifying the likely potential causes of a patient's disease or condition and weighing the risks of harms and benefits of additional testing or treatment, physicians then try to determine the most appropriate approach—testing, medication, or surgery, for example.¹⁰

Courts have used the term differential etiology interchangeably with differential diagnosis.¹¹ In medicine, etiology starts with the study of causation in disease,¹² but differential etiology is a legal invention not used by physicians. In general, both differential etiology and differential diagnosis are concerned with establishing or refuting causation between an external cause and a plaintiff's condition. Depending on the type of case and the legal standard, a medical expert may testify regarding specific causation, general causation, or both. Keep in mind that physicians typically make decisions about treatment and make

omitted) (stating expert's incomplete execution of differential diagnosis procedure rendered expert testimony unsatisfactory for *Daubert* standard); *Lash v. Hollis*, 525 F.3d 636, 640 (8th Cir. 2008) ("Further, even if the treating physician had specifically opined that the Taser discharges caused rhabdomyolysis in Lash Sr., the physician offered no explanation of a differential diagnosis or other scientific methodology tending to show that the Taser shocks were a more likely cause than the myriad other possible causes suggested by the evidence.") (finding lack of expert testimony with differential diagnosis enough to render evidence insufficient for jury to find causation in personal-injury suit); *Feit v. Great West Life & Annuity Ins. Co.*, 271 F. App'x 246, 254 (3d Cir. 2008) ("However, although this Court generally recognizes differential diagnosis as a reliable methodology the differential diagnosis must be properly performed in order to be reliable. To properly perform a differential diagnosis, an expert must perform two steps: (1) 'Rule in' all possible causes of Dr. Feit's death and (2) 'Rule out' causes through a process of elimination whereby the last remaining potential cause is deemed the most likely cause of death.") (citations omitted) (ruling that district court was not in error for excluding expert medical testimony that relied on an improperly performed differential diagnosis); *Saldana v. Delta Airlines, Inc.*, 1:19 CIV. 027 (CAK), 2021 WL 4710811 (D.V.I. Oct. 8, 2021) (quoting *Feit*, also considering that differential diagnosis had been properly performed in the case to form the opinion that the plaintiff had suffered a heart attack as a result of hurrying to catch the connecting flight).

9. *Stedman's Medical Dictionary* 531 (28th ed. 2006) (defining differential diagnosis as "the determination of which of two or more diseases with similar symptoms is the one from which the patient is suffering, by a systematic comparison and contrasting of the clinical findings.").

10. *The Concise Dictionary of Medical-Legal Terms* 36 (1998) (definition of differential diagnosis).

11. See *Proctor v. Fluor Enters., Inc.*, 494 F.3d 1337 (11th Cir. 2007) (testifying medical expert employed differential etiology to reach a conclusion regarding the cause of plaintiff's stroke). But see *McClain v. Metabolife Int'l, Inc.*, 401 F.3d 1233, 1252 (11th Cir. 2005) (distinguishing differential diagnosis from differential etiology, with the former closer to the medical definition and the latter employed as a technique to determine external causation); *Brown v. Burlington N. Santa Fe Ry. Co.*, 765 F.3d 765 (7th Cir. 2014) (indicating that differential diagnosis focuses on the identity of an ailment, while differential etiology focuses on the cause of the ailment).

12. *Stedman's Medical Dictionary* 675 (28th ed. 2006) (defining etiology as "the science and study of the causes of disease and their mode of operation"). For a discussion of the term etiology in epidemiology studies, see Steve C. Gold et al., *Reference Guide on Epidemiology*, section titled "Specific Causation," in this manual.

inferences about the possible diseases (different diagnoses) that may cause the presenting symptoms or signs to arrive at an appropriate treatment based on potential diagnoses when the diagnosis alters the optimal treatment. For example, stroke symptoms (the disease) can be caused by a variety of etiologies, mechanisms by which oxygen to the brain is disturbed (e.g., a burst brain aneurysm, a blood clot or an atherosclerotic plaque traveling to the brain, or other causes), thereby motivating further diagnostic testing to identify the etiology and treat appropriately. Physicians also use inference to help narrow potential diseases by identifying associated symptoms and risk factors—for example, chest pain due to coronary artery disease with a history of high blood pressure, high cholesterol, smoking, or occurring with exertion. The focus, however, is on diagnosis, typically with residual uncertainty and a weighing of the likelihood of that disease and the benefits and harms of treatment in a patient’s particular health context—age, sex, gender, past medical history, health behaviors and exposures (e.g., smoking, alcohol, occupational hazards, toxins), and family history—sometimes requiring revisiting the patient’s response to treatment over time and possibly ordering additional tests or treatments in an iterative approach. Most treating physicians will less commonly be familiar with differential etiology. General causation refers to whether the plaintiff’s injury could have been caused by the defendant or a product produced by the defendant, while specific causation is established only when the defendant’s actions or product actually caused the harm.¹³ An opinion by a testifying physician may be offered in support of both types of causation.¹⁴

Courts also refer to medical certainty or probability in ways that differ from their use in medicine. The standards of “reasonable medical certainty”¹⁵ and “reasonable medical probability” are also terms of art in the law that have no analog for a practicing physician.¹⁶ As detailed in the “Medical Decision-Making”

13. See *Amorgianos v. Nat’l R.R. Passenger Corp.*, 303 F.3d 256, 268 (2d Cir. 2002).

14. See, e.g., *Ruggiero v. Warner-Lambert Co.*, 424 F.3d 249 (2d Cir. 2005) (excluding testifying expert’s differential diagnosis in support of a theory of general causation because it was not supported by sufficient evidence); *Milward v. Rust-Oleum Corp.*, 820 F.3d 469 (1st Cir. 2016) (establishing that the plaintiffs had the burden of establishing, through expert testimony, general and specific causation, and analyzing the admissibility of an expert witness to establish the latter).

15. See, e.g., the official reporter’s note in *Restatement (Third) of Torts: Liability for Physical and Emotional Harm* § 28 cmt. e (Am. L. Inst. 2010), which exhaustively reviews the case law and law-review articles on this topic and finds that at least thirty-eight states have concluded that expert testimony otherwise admissible need not be expressed to a “reasonable degree of medical or scientific certainty.”

16. See, e.g., *Dallas v. Burlington N., Inc.*, 689 P.2d 273, 277 (Mont. 1984) (“[R]easonable medical certainty’ standard; the term is not well understood by the medical profession. Little, if anything, is ‘certain’ in science. The term was adopted in law to assure that testimony received by the fact finder was not merely conjectural but rather was sufficiently probative to be reliable.”); *Laue v. Voyles*, CV 13-31-BU-CSO, 2014 WL 12888669 (D. Mont. Apr. 18, 2014) (quoting the same passage from *Dallas*). This reference guide will not probe substantive legal standards in any detail, but there are substantive differences in admissibility standards for medical evidence between federal

section below, diagnostic reasoning and medical evidence are aimed at recommending the best therapeutic option for a patient. Although most courts have interpreted “reasonable medical certainty” to mean a preponderance of the evidence,¹⁷ physicians often work with multiple hypotheses while diagnosing and treating a patient without any “standard of proof” to satisfy.

Courts also use standard of care as a legal term that varies from state to state. In healthcare, medical-specialty societies use terms such as standards (high degree of certainty), guidelines (varying degrees of certainty), and consensus statements (expert opinion where controversy exists).¹⁸ Per the Institute of Medicine (IOM), “*Practice Guidelines* are systematically developed statements to assist practitioner and patient decisions about appropriate health care for specific clinical circumstances”; “*Medical Review Criteria* are systematically developed statements that can be used to assess the appropriateness of specific health care decisions, services, and outcomes”; “*Standards of Quality* are authoritative statements of (1) minimum levels of acceptable performance or results, (2) excellent levels of performance or results, or (3) the range of acceptable performance or results”; and “*Performance Measures* are methods or instruments to estimate or monitor the extent to which the actions of a health care practitioner or provider conform to practice guidelines, medical review criteria, or standards of quality.”¹⁹ In 2011, the Institute of Medicine updated the definition: “Clinical practice guidelines are statements that include recommendations intended to optimize

and state courts. See Robin Kundis Craig, *When Daubert Gets Erie: Medical Certainty and Medical Expert Testimony in Federal Court*, 77 Denv. U. L. Rev. 69 (1999).

17. See, e.g., *Sharpe v. United States*, 230 F.R.D. 452, 460 (E.D. Va. 2005) (“It is not enough for the plaintiff’s expert to testify that the defendant’s negligence might or may have caused the injury on which the plaintiff bases her claim. The expert must establish that the defendant’s negligence was ‘more likely’ or ‘more probably’ the cause of the plaintiff’s injury . . .”); *W.C. v. Sec’y of Health & Hum. Servs.*, 100 Fed. Cl. 440 (2011), *aff’d*, 704 F.3d 1352 (Fed. Cir. 2013) (finding that claimant had not established by preponderant evidence a medical theory causally connecting his significantly worsened multiple sclerosis to his flu vaccination).

18. In *Adams v. Lab. Corp. of Am.*, 760 F.3d 1322 (11th Cir. 2014), the issue was whether a litigation review of Pap smear slides needed to be done “blinded,” with the plaintiff’s specimen reviewed alongside others and with the expert not knowing which specimen was being challenged in the lawsuit. The court pointed out that there were plenty of times in pathology when retrospective reviews were done without such an elaborate “blinded” process, which seemed more suited to making it difficult to pursue litigation. As the Eleventh Circuit found, “As far as we are aware, this is the first time that an industry group has promulgated a set of guidelines that attempts to define and limit the evidence courts should accept when the group’s members are sued. The members of the CAP and ASC have a substantial interest in making it more difficult for plaintiffs to sue based on alleged negligence in their Pap smear screening, and their guidelines do just that.” *Id.* at 1331.

19. Comm. to Advise Pub. Health Serv. on Clinical Prac. Guidelines, Inst. Med., *Clinical Practice Guidelines: Directions for a New Program* 8 (Marilyn J. Field & Kathleen N. Lohr eds., 1990), <https://doi.org/10.17226/1626>. [hereinafter 1990 CAPHSCPG Report]. The Institute of Medicine is now the National Academy of Medicine.

patient care that are informed by a systematic review of evidence and an assessment of the benefits and harms of alternative care options.”²⁰

Statutes and administrative regulations may also contain terms that are borrowed, often imperfectly, from the medical profession. In these cases, the court may need to examine the intent of the legislature and the term’s usage in the medical profession. If no intent is apparent, the court may need to determine whether the medical definition is the most appropriate one to apply to the statutory language. Whether the language is a medical term of art or a question of law will often dictate the admissibility and weight of evidence.²¹

Testimony by Physicians

Federal Rule of Civil Procedure 26(a)(2) requires various disclosures regarding expert witnesses. In 2010, the expert disclosure requirements were expanded to include disclosure of anticipated testimony by non-retained experts, including treating physicians, even if those opinions were confined entirely to opinions formed during treatment. Law professor Steven Gensler summarizes the disclosure requirements in this way:

In summary, the disclosures that must be made for a treating physician depend on the nature of the testimony he or she will give. Unless the treating physician is going to be limited to testifying about facts in a lay person capacity, the physician must be disclosed as an expert and must provide either the summary disclosures or an expert report. Whether the treating physician must file a written report or is subject only to summary disclosures depends on the role of the expert. If the treating physician’s expert opinions stay within the scope of treatment and diagnosis, then the physician would not be considered “retained” to provide expert testimony and only summary disclosures would be needed. But if a treating physician is going to offer opinions formed outside the course of treatment and diagnosis, then as to those further opinions the physician is being used in a “retained expert” role and the Rule 26(a)(2)(B)’s report requirement will apply to the extent of that further testimony.²²

Thus, the level of disclosure required by Rule 26 turns on whether the treating physician confines his or her testimony to facts and opinions in the context of

20. Comm. on Standards for Developing Trustworthy Clinical Prac. Guidelines, Inst. Med., Clinical Practice Guidelines We Can Trust 4 (Robin Graham et al. eds., 2011), <https://doi.org/10.17226/13058>. [hereinafter 2011 CSDTCGP Report].

21. See, e.g., *Coleman v. Workers’ Comp. Appeal Bd. (Ind. Hosp.)*, 842 A.2d 349 (Pa. 2004) (holding that since the legislature did not define the medical term *physical examination*, the common usage of the term is more appropriate than the strict medical definition).

22. Steven S. Gensler, 1 Federal Rules of Civil Procedure: Rules and Commentary, Rule 26 (2023).

determining the treatment, or ventures into opinions formed apart from treatment. In practice, parties rarely object on Rule 702 grounds to a doctor's testimony confined to treatment, even if it includes opinions formed during treatment. But when a treating clinician gives opinions beyond those formed during treatment, such as an opinion on future costs of treatment, *Daubert* objections are more common.

Although testifying medical experts often will need to tailor their opinions in a way that conforms to the legal standard of causation, the treating physicians generally are more concerned about accurate diagnosis and treatment and are less familiar with differential etiology. As the section titled "Medical Decision-Making" below will demonstrate, when analyzing the patient's symptoms and making a judgment based on the available medical evidence, a treating physician may not expressly identify a "proximate cause" or "substantial factor." For example, in order to recommend treatment, a physician does not necessarily need to determine whether a patient's lung ailment was more likely the result of a long history of tobacco use or prolonged exposure to asbestos if the optimal treatment is the same. In contrast, when testifying as an expert in a case in which an employee with a long history of tobacco use is suing his employer for possible injuries from asbestos exposure in the workplace, physicians may need to make judgments regarding the likelihood that either tobacco or asbestos—or both—could have contributed to the injury.²³

Physicians may be asked to testify about patients they have never examined or from whom they have never taken a medical history and make assessments of the proximate cause, increased risk of injury, or likely future injuries.²⁴ A doctor may even need to make medical judgments about a deceased litigant.²⁵ Testifying in all such cases requires making judgments that physicians do not ordinarily make in their profession and adapting their opinion to take into account the appropriate legal standard.²⁶

23. Physicians will testify as experts in cases in which the plaintiff's condition may be the result of multiple causes. In these cases, the divergence between medical reasoning and legal reasoning is very apparent. See, e.g., *Tompkin v. Philip Morris USA, Inc.*, 362 F.3d 882 (6th Cir. 2004) (affirming district court's conclusion that testimony offered by defendant's expert regarding the decedent's work-related asbestos exposure was not prejudicial in a suit against a tobacco company on behalf of plaintiff's deceased husband); *Mobil Oil Corp. v. Bailey*, 187 S.W.3d 265 (Tex. Ct. App. 2006) (involving claims from a worker who had a long history of tobacco use that exposure to asbestos increased his risk of cancer).

24. Visual and verbal information about patients suspected of having acute heart disease influenced a physician's estimates of the likelihood of disease by 35% (from a mean of 51%, with an absolute change of 18%). See C.P. Friedman et al., *Visual Information and the Diagnosis of Chest Pain*, 69 Acad. Med. S28–S30 (1994), <https://doi.org/10.1097/00001888-199410000-00032>.

25. See, e.g., *Tompkin*, 362 F.3d 882.

26. More than thirty medical specialty societies have issued statements about ethical requirements for their members to testify in court. That also deserves discussion. A prominent example is the American Academy of Orthopedic Surgeons (AAOS), which has an "expert witness affirmation statement," available at <https://perma.cc/H7TA-LC9T>.

Medical Expertise and Admissibility

As the “Medical Testimony Introduction” section above suggested, the goal that guides the physician—recommending the best therapeutic options for the patient—means that diagnostic reasoning involves probabilistic judgments concerning several working hypotheses, often considered simultaneously. Legal standards for physician testimony include Federal Rule of Civil Procedure 26(a)(2) requiring various disclosures regarding expert witnesses. The 1993 Advisory Committee Note to Rule 26—when expert disclosure requirements in the absence of a discovery request were added to Rule 26—specifically states that experts as used in Rule 26 are those who testify under Rule 702. Before 2010, the only expert disclosures required of parties were the detailed reports required for experts “retained or specially employed” to give testimony in the case.²⁷ This usually did not include treating physicians, who were testifying because they provided treatment and not because they were hired by a party. In 2010, the expert disclosure requirements were expanded to include non-retained experts who testify at trial.²⁸ This was specifically done to address treating-physician testimony, which normally was not subject to the report requirement but often was pivotal at trial.

Therefore, treating doctors who do more than recite basic facts are giving testimony subject to Rule 702, but the admissibility of their testimony is typically more of an issue—and *Daubert* challenges are more common—when they provide opinions not formed during the course of their treatment. The “Medical Care” and “Medical Decision-Making” sections below explain in great detail the practice of medicine, including medical education, the structure of health-care, and, most importantly, the methods that physicians use to diagnose and treat their patients. Special attention is given to the physician-patient relationship and to the types of evidence that physicians use to make medical judgments.

Medical Care

Medical Education and Training

Medical School

The Association of American Medical Colleges (AAMC) oversees accredited U.S. medical schools, as well as 197 Canadian medical schools.²⁹ The Liaison Committee on Medical Education performs the accreditation for AAMC and

27. Fed. R. Civ. P. 26(a)(2)(B).

28. Fed. R. Civ. P. 26(a)(2)(A), (C).

29. Ass’n Am. Med. Colls., Membership, <https://perma.cc/49NB-MLXW>.

assesses the quality of post-secondary education by determining whether each institution or program meets established standards for function, structure, and performance. All physicians who wish to be licensed must pass the U.S. Medical Licensing Examination Steps 1, 2, and 3.³⁰

In the United States, the bulk of physicians are allopathic medical doctors (M.D.s) while 11% are doctors of osteopathy (D.O.s).³¹ The Commission on Osteopathic College Accreditation accredits osteopathic medical schools. Training is similar to that for allopathic physicians but with “special training in the musculoskeletal system, your body’s interconnected system of nerves, muscles and bones.”³² About 25% of current U.S. physicians are foreign medical graduates that

30. U.S. Medical Licensing Examination, *General Common Questions*, <https://perma.cc/P67R-EESM>.

Rule 702 requires an expert to be qualified by “knowledge, skill, experience, training, or education.” *Planned Parenthood Cincinnati Region v. Taft*, 444 F.3d 502, 515 (6th Cir. 2006) (“The State has not appealed the district court’s order refusing to recognize Dr. Crockett as an expert in the critical review of medical literature. Although that order has not been placed before us, the only reason the district court gave for her ruling was that Dr. Crockett did not have any specific training in the critical review of medical literature beyond the training incorporated in her general medical school and residency training. This ruling ignored Dr. Crockett’s testimony that her residency program at Georgetown University put particular emphasis on training residents in the critical review of medical literature, that she had taught classes on the subject, that she had done extensive reading and self-education on the subject, and that she had critically reviewed medical literature for the FDA. If these qualifications are not sufficient to demonstrate expertise, this court is hard-pressed to imagine what qualifications would suffice.”); *Davis v. Houston Cnty., Ala. Bd. of Educ.*, No. 1:06-CV-953-MEF, 2008 WL 410619, at *4 (M.D. Ala. Feb. 13, 2008), *aff’d sub nom. Davis v. Houston Cnty., Ala. Bd. of Educ.*, 291 F. App’x 251 (11th Cir. 2008) (“The Board has moved to exclude all evidence of Freet’s opinions and conclusions related to the cause of Joshua Davis’s behavior at the football game contained in his deposition as well as Freet’s letter to Malcolm Newman. The Board argues that Freet is not qualified to give expert testimony, and that Plaintiff failed to comply with Fed. R. Civ. 26(a)(2)(B) by not providing a report of Freet’s testimony that includes all of the information required by Rule 26(a)(2)(B). . . . In order to consider Freet’s expert opinions, this Court must find that Freet meets the requirements of Fed. R. Evid. 702. . . . Freet is not a medical doctor and never attended medical school. The only evidence of Freet’s qualifications are: approximately five years working for the Department of Veterans Affairs in the vocational rehabilitation program, followed by approximately seven years working in private practice as a ‘licensed professional counselor.’ There is no evidence in the record of Freet’s educational background, or any details of the exact nature of Freet’s work experience.”); *Therrien v. Town of Jay*, 489 F. Supp. 2d 116, 117 (D. Me. 2007) (“Citing *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 509 U.S. 579, 113 S. Ct. 2786, 125 L. Ed. 2d 469 (1993) and Rule 702 of the Federal Rules of Evidence, Officer Gould’s first objection is that Dr. Harding does not possess sufficient expertise to express expert opinions about ‘the mechanism and timing of Plaintiff’s injuries.’ This objection is not well taken. Dr. Harding was graduated from Dartmouth College and Georgetown Medical School; he completed a residency in internal medicine, is board certified in internal medicine, and has been licensed to practice medicine in the state of Maine since 1978.”).

31. Am. Osteopathic Ass’n, *What Is a DO?*, <https://perma.cc/Q2AT-5SPX>.

32. *Id.*

include both U.S. citizens and foreign nationals.³³ Because educational standards and curricula outside the United States and Canada vary, the Education Commission for Foreign Medical Graduates administers a certification exam to determine eligibility for Accreditation Council for Graduate Medical Education (ACGME) accredited U.S. residency and fellowship programs.³⁴

Postgraduate Training

Most medical school graduates seek additional training through residency programs in their chosen specialty.³⁵ Residencies range from three to seven years at teaching hospitals and academic medical centers. In training, residents care for patients while being supervised mostly by board-certified physicians. They also have opportunities to pursue educational and research activities.³⁶ Physician licensure in many states requires the completion of a residency program accredited by the ACGME, which is responsible for all allopathic and osteopathic graduate medical education programs, including over 13,000 residency and fellowship programs in 182 specialties and subspecialties.³⁷ After residency, eligible physicians can take their board certification examinations (see below), and some opt for additional subspecialty fellowship training.

Licensure and Credentialing

Medical practice acts defining the practice of medicine and delegating enforcement to state medical boards exist for each of the fifty states, the District of Columbia, and the U.S. territories. State medical boards award medical licenses, investigate complaints, discipline physicians who violate the law, and evaluate and rehabilitate physicians.

The hospital credentialing process defines physicians' scope of practice and hospital privileges, that is, the clinical inpatient and outpatient services and procedures they may provide. The process involves verifying medical education, postgraduate training, board certification, professional experience, state licensure, prior credentialing outcomes, medical-board actions, malpractice, and adverse clinical

33. Ass'n of Am. Med. Colls., *2023 U.S. Physician Workforce Data Dashboard*, <https://perma.cc/J59U-Q3VZ>.

34. Educ. Comm'n for Foreign Med. Graduates, *About ECFMG*, <https://perma.cc/C3N8-9HHX>.

35. See *Brown v. Hamot Med. Ctr.*, Civil Action No. 05–32E, 2008 WL 55999 (W.D. Pa. Jan. 3, 2008), *aff'd*, 323 F. App'x 140 (3d Cir. 2009).

36. See *Planned Parenthood Cincinnati Region v. Taft*, 444 F.3d 502, 515 (6th Cir. 2006).

37. Accreditation Council for Graduate Med. Educ., *About the ACGME*, <https://perma.cc/P8XQ-8ZX2>.

events. Re-credentialing involves an assessment of a physician's professional or technical competence and performance by evaluating and monitoring the quality of patient care. Licensing and credentialing is identical for allopathic medical doctors and doctors of osteopathy.

The American Board of Medical Specialties (ABMS) provides certification in twenty-four medical boards, forty specialties, and eighty-eight subspecialties³⁸ to maintain and evaluate standards in the profession and to help “physicians and specialists demonstrate their competence and professionalism throughout their careers.”³⁹ Although criteria vary by field, board eligibility requires completing medical school, passing the United States Medical Licensing Examination, finishing an appropriate residency, and evaluation generally with computer-based and, in some cases, oral examinations. ABMS Standards for Continuing Certification, effective January 1, 2024, seek to promote integrated, specialty-specific programs by member boards to support an individual physician's or medical specialist's (i.e., diplomate's) continuing professional development.⁴⁰ In some cases, specialty organizations have developed their own certification process outside of the ABMS (e.g., the American Board of Bariatric Medicine).⁴¹ The American Osteopathic Association (AOA) certifies osteopathic physicians in sixteen specialty boards, issues certificates in twenty-seven primary specialties and forty-eight subspecialties,⁴² and has a continuous certification process.⁴³

Continuing Medical Education

For relicensure, state medical boards require continuing medical education so that physicians can acquire new knowledge and maintain clinical competence. The Accreditation Council for Continuing Medical Education (ACCME) identifies, develops, and promotes quality standards for organizations providing continuing

38. Am. Bd. Med. Specialties, <https://perma.cc/4DSJ-48HJ>.

39. Although specialization is a hallmark of modern medical practice, courts have not always required that medical testimony come from a specialist. *See, e.g., Gaydar v. Sociedad Instituto Gineco-Quirurgico y Planificacion Familiar*, 345 F.3d 15, 24–25 (1st Cir. 2003) (“The proffered expert physician need not be a specialist in a particular medical discipline to render expert testimony relating to that discipline.”); *Keller v. Feasterville Fam. Health Care Ctr.*, 557 F. Supp. 2d 671 (E.D. Pa. 2008) (holding that the expert opinion that patient did not have Alzheimer's disease was admissible in medical malpractice action, even though treating physicians were internal and family-care physicians, and expert was board certified in family medicine, but never received any specialized education, training, or experience in neuropathology, neurodegenerative diseases, or Alzheimer's disease).

40. Am. Bd. Med. Specialties, Standards for Continuing Certification, <https://perma.cc/5SQ2-UMWP>.

41. Am. Bd. Obesity Med., Certifying Physicians in the Treatment of Obesity, <https://perma.cc/7HUX-ASF6>.

42. Am. Osteopathic Ass'n, AOA Board Certification, <https://perma.cc/N5ZQ-88DY>.

43. Am. Osteopathic Ass'n, Osteopathic Continuous Certification, <https://perma.cc/S4E5-EF4F>.

medical education for physicians⁴⁴ to help assure physicians, state medical boards, medical societies, state legislatures, continuing medical education providers, and the public that the education meets certain quality standards. For osteopathic physicians, the AOA Board of Trustees also oversees accreditation for osteopathic continuing medical education sponsors through the Council on Continuing Medical Education (CCME).⁴⁵

Organization of Medical Care

The delivery of healthcare in the United States is highly decentralized and fragmented.⁴⁶ Healthcare is provided through clinics, hospitals, managed-care organizations, medical groups, multispecialty clinics, integrated delivery systems, specialty standalone hospitals, imaging facilities, skilled-nursing facilities, rehabilitation hospitals, emergency departments, and pharmacy-based and other walk-in clinics. Transitions in care settings involve multiple handoffs among different healthcare professionals and care providers with the need for accurate, timely, and complete transfer of information. As hospitals increasingly belong to a network or system, physicians too have shifted toward larger practices. For the first time, fewer than half of patient-care physicians were in private practice in 2020; therefore, “[n]o single practice type, ownership structure, or size can or should be considered the typical physician practice.”⁴⁷ The Covid-19 pandemic has accelerated practice changes including telemedicine visits, home hospitalizations, and ambulatory surgery and procedure centers with their own specialty organizations, accreditation bodies, and state regulatory oversight, besides possible federal certification.

Per W. Edwards Deming, “Every system is perfectly designed to get the result that it does.” As a consequence, the U.S. healthcare system yields a life expectancy 5.9 years shorter than the average of comparable countries in 2021, yet it spends nearly twice as much per capita.⁴⁸ Driven by social and economic inequities, the healthcare system has substantial disparities by race/ethnicity, but also by “socioeconomic status, age, geography, language, gender, disability

44. Accreditation Council for Continuing Med. Educ., About Accreditation, <https://perma.cc/P3UF-QTKT>.

45. Am. Osteopathic Ass’n, CME, <https://perma.cc/6RBB-765P>.

46. Comm. on Quality Health Care Am., Inst. Med., *Crossing the Quality Chasm: A New Health System for the 21st Century* (2001) [hereinafter 2001 CQHCA Report], <https://doi.org/10.17226/10027>.

47. Carol K. Kane, Am. Med. Ass’n, *Recent Changes in Physician Practice Arrangements: Private Practice Dropped to Less Than 50 Percent of Physicians in 2020*, Policy Research Perspectives, <https://perma.cc/VQD4-RXX9>.

48. Shameek Rakshit, Matthew McGough & Krutika Amin, Peterson-KFF Health System Tracker, *How Does U.S. Life Expectancy Compare to Other Countries?* (Jan. 30, 2024), <https://perma.cc/J8T2-P6Q9>.

status, citizenship status, and sexual identity and orientation.” For example, “[b]lack infants in America are now more than twice as likely to die as white infants, a disparity that is wider than it was in 1850, 15 years before the end of slavery.”⁴⁹

Since a Harvard Medical Practice study found a high incidence of adverse events in hospitalizations and highly publicized errors (fatal medication overdoses and amputation of the limb on the wrong side), there have been safety concerns about the organization of medical care. The Harvard study indicated that adverse events occurred in 3.7% of hospitalizations and that 27.6% of adverse events were due to negligence.⁵⁰ A report by the IOM estimated that negligence errors resulted in as many as 98,000 deaths in patients hospitalized annually.^{51,52} The report further highlighted system issues leading to safety events: “The decentralized and fragmented nature of the healthcare delivery system (some would say ‘nonsystem’) also contributes to unsafe conditions for patients and serves as an impediment to efforts to improve safety.” While recognizing that “not all errors result in harm,” the report defined safety as “freedom from accidental injury” and specified two types of error: “the failure of a planned action to be completed as intended or the use of a wrong plan to achieve an aim.”⁵³

The IOM recommended the development of a learning healthcare–delivery system—“a system that both prevents errors and learns from them when they occur. The development of such a system requires, first, a commitment by all stakeholders to a culture of safety and, second, improved information systems.”⁵⁴

49. Nambdi Ndugga & Samantha Artiga, Peterson–KFF Health System Tracker, *Disparities in Health and Health Care: 5 Key Questions and Answers* (Apr. 21, 2023), <https://perma.cc/ZGU7-U9SB>; Linda Villarosa, *Why America's Black Mothers and Babies Are in a Life-or-Death Crisis*, N.Y. Times Mag., Apr. 11, 2018, <https://www.nytimes.com/2018/04/11/magazine/black-mothers-babies-death-maternal-mortality.html>.

50. A systematic review and meta-analysis identified 45 studies published between 2000 to 2018 in hospital settings from around the world and found that 10% of hospitalized patients experienced harm with half of those harms being preventable. See Maria Panagioti et al., *Prevalence, Severity, and Nature of Preventable Patient Harm Across Medical Care Settings: Systematic Review and Meta-Analysis*, 2019 B.M.J. 366 (Table 1, p.4 of PDF), <https://doi.org/10.1136/bmj.l4185>. See also Troyen A. Brennan et al., *Incidence of Adverse Events and Negligence in Hospitalized Patients—Results of the Harvard Medical Practice Study I*, 324 NEJM 370–76 (1991), <https://www.nejm.org/doi/full/10.1056/NEJM199102073240604>.

51. Comm. on Quality Health Care in Am., Inst. Med., *To Err Is Human: Building a Safer Health System* (Linda T. Kohn et al. eds., 2000) [hereinafter 2000 CQHCA Report], <https://doi.org/10.17226/9728>.

52. In 70 studies identified in a systematic review across global medical settings (e.g., primary care to intensive care or surgery) involving 337,025 patients, the overall pooled prevalence of harm was 12% with half of the harms being preventable. Among preventable harms, 12% were severe, e.g., resulting in permanent disability or death. See Panagioti et al., *supra* note 50.

53. 2000 CQHCA Report, *supra* note 51, at 3–4.

54. Comm. on Data Standards for Patient Safety, Inst. Med., *Patient Safety: Achieving a New Standard for Care 1* (Philip Aspden et al. eds. 2005), <https://doi.org/10.17226/10863>.

Government and nongovernment institutions have adopted as parts of their mission the assessment and promotion of safety at the healthcare-system level.

Besides patients and their families,⁵⁵ the work system for healthcare delivery and diagnosis includes organizational characteristics (culture, rules and procedures, leadership, and management), technologies and tools (health information technology and electronic health records), physical environment (layout, distractions, lighting, and noise), and external factors (payment, care-delivery system, legal system, and reporting environment).⁵⁶ Beyond physicians, medical delivery systems have increasingly incorporated allied health professions, including nurses, nurse practitioners, physician assistants, pharmacists, and therapists. The term *clinicians* will be used to be inclusive of this expanded diagnostic and care-delivery team.

Patient Care

Goals

The translated classical version of the Hippocratic Oath (c. 400 B.C.), “abstain from whatever is deleterious,” was abbreviated into “first do no harm,” or in Latin, *primum non nocere*. The IOM describes quality healthcare delivery as “[t]he degree to which health services for individuals and populations increase the likelihood of desired health outcomes and are consistent with current professional knowledge.” The six aims for improving the healthcare system assert that healthcare should be

- *Safe*—avoiding injuries to patients from the care that is intended to help them.
- *Effective*—providing services based on scientific knowledge to all who could benefit and refraining from providing services to those not likely to benefit (avoiding underuse and overuse, respectively).
- *Patient-centered*—providing care that is respectful of and responsive to individual patient preferences, needs, and values and ensuring that patient values guide all clinical decisions.
- *Timely*—reducing waits and sometimes harmful delays for both those who receive and those who give care.
- *Efficient*—avoiding waste, including waste of equipment, supplies, ideas, and energy.
- *Equitable*—providing care that does not vary in quality because of personal characteristics such as gender, ethnicity, geographic location, and socioeconomic status.⁵⁷

55. Kathryn M. McDonald et al., *The Patient Is In: Patient Involvement Strategies for Diagnostic Error Mitigation*, 22 B.M.J. Quality & Safety ii33 (2013), <https://doi.org/10.1136/bmjqs-2012-001623>.

56. Comm. on Diagnostic Error in Health Care, Nat'l Acads. Sci., Eng'g & Med., *Improving Diagnosis in Health Care* 34 (Erin P. Balogh et al. eds., 2015), <https://doi.org/10.17226/21794> [hereinafter 2015 CDEHC Report].

57. 2001 CQHC Report, *supra* note 46, at 5–6.

In *Crossing the Quality Chasm*, the IOM emphasized that care delivery should accommodate individual patient choices and preferences and be customized on the basis of a patient's needs and values.⁵⁸ The Charter on Medical Professionalism specifies three fundamental principles: (1) patient welfare or serving the interest of the patient, (2) patient autonomy or empowering patients to make informed decisions, and (3) social justice or fair distribution of healthcare resources.⁵⁹ When faced with limited resources, such as when a patient gets an intensive care bed during the Covid-19 pandemic, clinicians are faced with the dilemma of choosing between patient care and societal goals.⁶⁰

Diagnostic Process

The diagnostic problem begins with patients experiencing a health problem that leads to a healthcare system encounter (Figure 1).⁶¹ Initial information gathering then leads to information integration and interpretation in which a set of working diagnoses are formed and revised until sufficient information has been gathered in the clinical reasoning process. The sources of information include the clinical history and interview, physical examination, diagnostic testing, and possible referral and consultation.⁶²

The clinical history and interview identify the chief complaint as the specific symptom that led the patient to seek medical attention. The history of the present illness describes the onset and progression of symptoms over time and may include eliciting the presence or absence of symptoms as “pertinent positives or pertinent negatives” that increase or decrease the likelihood of potential diseases under consideration. The past medical history includes prior illnesses, hospitalizations, surgeries, current medications, drug allergies, and lifestyle habits (smoking, alcohol use, illicit drug exposure, dietary habits, and exercise habits). Family history captures illnesses diagnosed in related family members. Social history includes education, employment, and social relationships. These provide a context for the chief complaint. Finally, the review of systems is a comprehensive inquiry of symptoms from various organ systems.

58. *Id.* at 8.

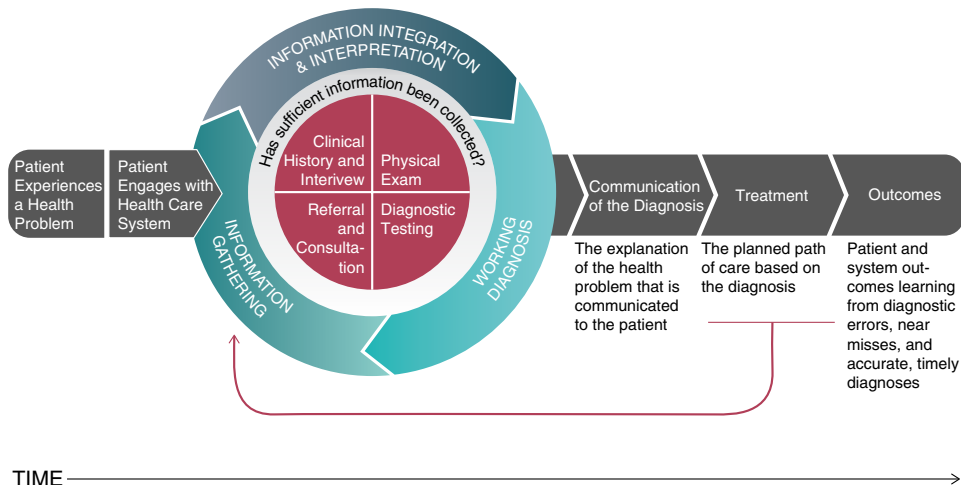
59. ABIM Found., ACP-ASIM Found. & Eur. Fed'n Internal Med., *Medical Professionalism in the New Millennium: A Physician Charter*, 136 *Annals Internal Med.* 243 (2002), <https://doi.org/10.7326/0003-4819-136-3-200202050-00012>.

60. *Medical Professionalism and the Parable of the Craft Guilds*, 147 *Annals Internal Med.* 809 (Harold C. Sox ed., 2007), <https://doi.org/10.7326/0003-4819-147-11-200712040-00015>.

61. See generally *Davoll v. Webb*, 194 F.3d 1116, 1138 (10th Cir. 1999) (“A treating physician is not considered an expert witness if he or she testifies about observations based on personal knowledge, including treatment of the party.”).

62. 2015 CDEHC Report, *supra* note 56, at 32–41.

Figure 1. The diagnostic process.



Source: From National Academies of Sciences, Engineering, and Medicine, *Improving Diagnosis in Health Care* 33 (National Academies Press 2015), <https://doi.org/10.17226/21794>.

Distinct from symptoms that are described by patients, physical-examination findings are considered as signs. Directed physical examination refers to inspecting only the relevant organ systems that may be causing the symptoms.

Based on the set of working diagnoses, diagnostic testing may be ordered to confirm or rule out possible diagnoses. Diagnostic testing is distinguished from surveillance testing or screening because the patient has a symptom whose cause the diagnostic team is seeking to identify for treatment. Depending on residual diagnostic uncertainty, patients may be referred for further consultation with specialists or subspecialists for their diagnostic or therapeutic opinion.

Patients also may seek care to monitor chronic conditions. This places an emphasis on collaborative and continuous care that involves patients, their families, clinicians, long-term care goals and plans, self-management training, and support⁶³ with organizational needs that differ substantially from those necessary for acute episodic complaints. The clinical history and physical examination involve assessing whether the symptoms or signs of that condition have progressed, improved, or stabilized. Surveillance may involve repeating tests at some recommended frequency for a known health risk—for example, a colonoscopy in three years for someone found to have premalignant polyps previously, instead of ten years for someone with a negative colonoscopy.

63. 2001 CQHCA Report, *supra* note 46, at 27.

Patients may seek preventive health visits. Screening involves testing asymptomatic, otherwise-healthy patients to find a disease earlier, presumably when it may be more treatable (e.g., cancer) or preventable (e.g., preventing stroke by treating hypertension, a risk factor for stroke). Screening for a population requires that (1) the condition affects quality and length of life in the population; and (2) has a sufficiently high incidence or prevalence to justify any risk of harms associated with the test; (3) a preventive or early treatment is available; (4) an asymptomatic period for early detection is present; (5) an accurate, acceptable, and affordable screening test exists; and (6) as applied to a population, screening benefits should exceed harms. Screening for disease in asymptomatic, otherwise healthy patients has become widely accepted and promulgated.⁶⁴ In contrast to diagnostic testing prompted by a complaint from a patient, screening involves apparently healthy individuals without any complaint,⁶⁵ so “every adverse outcome of screening is iatrogenic and entirely preventable” by not screening, including overdiagnosis and overtreatment (diagnoses that never would have been made, e.g., never cause symptoms or mortality resulting in over treatment with potential adverse effects).⁶⁶

Medical Decision-Making

Uncertainty in defining when a disease is present makes diagnosis, and therefore treatment decisions, difficult: (1) the difference between normal and abnormal is not always well demarcated; (2) many diseases do not progress with certainty (e.g., progression of lobular carcinoma in situ of the breast to invasive breast cancer occurs less than 50% of the time), but rather increase the risk of a poor outcome (e.g., hypertension raises the risk of developing heart disease or stroke); and (3) symptoms, signs, and findings for one disease overlap with others.⁶⁷ Variation also exists in the ability of physicians to elicit particular symptoms (e.g., in a group of patients interviewed by many physicians, 23% to 40% of the physicians reported cough as being present), observe signs (e.g., only 53% of physicians detected cyanosis—a blue or purple discoloration of the skin resulting from lack of oxygen—when present), or interpret tests (e.g., only 51% of pathologists agreed with each other when examining Pap smear slides with abnormal cells taken from

64. Lisa M. Schwartz et al., *Enthusiasm for Cancer Screening in the United States*, 291 JAMA 71 (2004), <https://doi.org/10.1001/jama.291.1.71>.

65. David A. Grimes & Kenneth F. Schulz, *Uses and Abuses of Screening Tests*, 359 Lancet 881, 881 (2002), [https://doi.org/10.1016/S0140-6736\(02\)07948-5](https://doi.org/10.1016/S0140-6736(02)07948-5).

66. *Id.* at 881; William C. Black, *Overdiagnosis: An Underrecognized Cause of Confusion and Harm in Cancer Screening*, 92J. Nat’l Cancer Inst. 1280, 1280 (2000), <https://doi.org/10.1093/jnci/92.16.1280>.

67. David M. Eddy, *Variations in Physician Practice: The Role of Uncertainty*, 3 Health Affs. 74, 75–76 (1984), <https://doi.org/10.1377/hlthaff.3.2.74>.

a woman's cervix to look for signs of cervical cancer).⁶⁸ And prognosis (response to disease or treatment) with alternative therapies is in many cases uncertain. A report by the Royal College of Physicians puts it this way:

The practice of medicine is distinguished by the need for judgement in the face of uncertainty. Doctors take responsibility for these judgements and their consequences. A doctor's up-to-date knowledge and skill provide the explicit scientific and often tacit experiential basis for such judgements. But because so much of medicine's unpredictability calls for wisdom as well as technical ability, doctors are vulnerable to the charge that their decisions are neither transparent nor accountable.⁶⁹

Note that a key piece of the diagnostic process in Figure 1 involves communication of the diagnosis and the care-path plan for the diagnosis.

Clinical-Reasoning Process

Studies of clinical problem solving suggest that physicians employ combinations of two diagnostic approaches ranging from hypothetico-deductive reasoning (deliberative and analytical, or thinking slow) to heuristic pattern recognition (quick and intuitive, or thinking fast).⁷⁰

Hypothetico-Deductive Reasoning

In the hypothetico-deductive approach, based on initial information such as age, gender, and chief complaint, clinicians begin to generate an initial list of potential diseases (hypothesis generation) based on their potential to explain the patient's observed signs and symptoms (Figure 2)⁷¹ within the known capacity

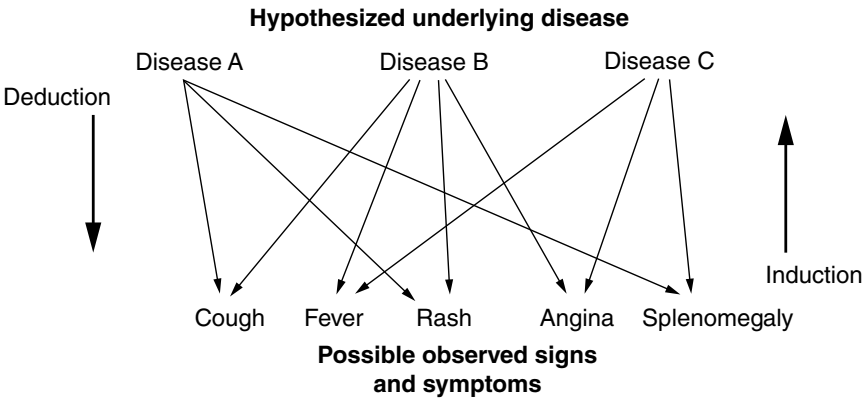
68. *Id.* at 77–78.

69. Royal Coll. Physicians, *Doctors in Society: Medical Professionalism in a Changing World* xi, <https://perma.cc/82RF-6GX6>.

70. Jerome P. Kassirer et al., *Learning Clinical Reasoning* 5–7, 56–88 (2d ed. 2010); Arthur S. Elstein & Alan Schwartz, *Clinical Problem Solving and Diagnostic Decision Making: Selective Review of the Cognitive Literature*, 324 B.M.J. 729, 729–30 (2002), <https://doi.org/10.1136/bmj.324.7339.729>; Jerome P. Kassirer & G. Anthony Gorry, *Clinical Problem Solving: A Behavioral Analysis*, 89 *Annals Internal Med.* 245 (1978), <https://doi.org/10.7326/0003-4819-89-2-245>; Geoffrey Norman, *Research in Clinical Reasoning: Past History and Current Trends*, 39 *Med. Educ.* 418 (2005), <https://doi.org/10.1111/j.1365-2929.2005.02127.x>.

71. Steven N. Goodman, *Toward Evidence-Based Medical Statistics. 1: The P Value Fallacy*, 130 *Annals Internal Med.* 995, 996 (1999), <https://doi.org/10.7326/0003-4819-130-12-199906150-00008>.

Figure 2. Medical deductive and inductive inference.



Source: Used with permission of the American College of Physicians, from Steven N. Goodman, *Toward Evidence-Based Medical Statistics. 1: The P Value Fallacy*, 130 *Annals Internal Med.* 995, 996 (1999), <https://doi.org/10.7326/0003-4819-130-12-199906150-00008>; permission conveyed through Copyright Clearance Center, Inc.

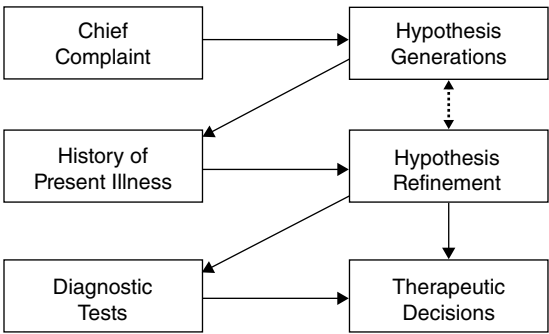
limitations in human short-term memory.⁷² This initial working list of diagnoses provides a context used to evaluate subsequent information. Based on their disease knowledge, clinicians expect the presence or absence of certain symptoms, risk factors, disease course, signs, or test results for each diagnosis (deductive inference).

As further information emerges (Figure 3), those data are evaluated and interpreted for their consistency with the working list of possibilities and whether those data would increase or decrease the likelihood of each possibility (hypothesis refinement). If the data are inconsistent, those diseases may be dropped and other diagnostic possibilities may be considered (hypothesis modification). The information gathering continues iteratively over time, including possible referral to another clinician. The final cognitive step (diagnostic verification) involves testing the validity of the diagnosis for its coherency (consistency with predisposing risk factors, physiological mechanisms, and resulting manifestations), its adequacy (the ability to account for all normal and abnormal findings and the disease time course), and its parsimony (the simplest single explanation as opposed to requiring the simultaneous occurrence of two or more diseases to explain the findings).⁷³

72. Elstein & Schwartz, *supra* note 70, at 732; George A. Miller, *The Magical Number Seven Plus or Minus Two: Some Limits on Our Capacity for Processing Information*, 63 *Psych. Rev.* 81, 81 (1956), <https://doi.org/10.1037/h0043158>.

73. Kassirer et al., *supra* note 70, at 6, 8–16, 89–127.

Figure 3. Clinical-reasoning process from chief complaint to therapeutic decisions.



Heuristics: Intuitive Pattern Recognition

Alternatively, heuristics are quick, automatic “rules of thumb” or cognitive shortcuts. In such cases, pattern recognition leads to rapid recognition and a reflexive diagnosis with little cognitive effort.⁷⁴ For example, a black woman with large shadows from enlarged lymph nodes in her chest X-ray would trigger a presumptive diagnosis of sarcoidosis for many clinicians. The simplifying assumptions involved in heuristics, however, are subject to cognitive biases. For instance, episodic headache, sweating, and a rapid heartbeat form the classic triad seen in patients with a rare adrenal tumor known as a pheochromocytoma that also can cause hypertension. Physicians finding those three symptoms in a patient with hypertension may overestimate the patient’s likelihood of having pheochromocytoma based on representativeness bias, overestimating the likelihood of a less common disease just because case findings resemble those found in that disease.⁷⁵ Other cognitive errors include availability (overestimating the likelihood of memorable diseases in a subsequent patient because of vivid experience with a prior patient or media attention, and thus underestimating common or routine diseases) and

74. Stephen G. Pauker & John B. Wong, *How (Should) Physicians Think?: A Journey from Behavioral Economics to the Bedside*, 304 JAMA 1233, 1233–34 (2010), <https://doi.org/10.1001/jama.2010.1336>.

75. For additional discussion and definition of terms, see section titled “Diagnostic Reasoning” below. Applying Bayes’ rule, about 100 in 100,000 patients with hypertension have pheochromocytoma; this symptom triad occurs in 91% of patients with pheochromocytoma (sensitivity), and does not occur in 94% of those without pheochromocytoma (specificity), and so 6% of those without pheochromocytoma would have this symptom triad. Based on Bayes’ rule, 91 of the 100 individuals with pheochromocytoma (91% times 100) would have this triad, and 5,994 without a pheochromocytoma (6% times 99,900) will have this triad. Thus, among the 100,000 hypertensive patients, 6,085 will have the classic triad, suggesting the possibility of pheochromocytoma, but only 91 out of the 6,085 or 1.5%, will indeed have pheochromocytoma.

anchoring (insufficient adjustment of the initial likelihood of disease following a positive or negative test).⁷⁶

Clinical intuition refers to rapid, unconscious processes that select the pertinent findings from the multitude of available data.⁷⁷ Such expertise results from multiple exposures to patients with similar symptoms and their final diagnosis, is context sensitive, and cannot always be reduced to cause and effect.⁷⁸ Cognitive research into the development of expertise suggests two competing hypotheses. In instance- or exemplar-based memory, physicians store scripts or stories of prior recalled case examples—for example, visual information such as that in pathology, dermatology, or radiology—and match new cases to those stories. The alternative prototype memory hypothesis is based on a mental model of disease wherein experts store structured clinical “facts” to create abstractions of patients with the disease. These “prototypes” enable experts to link findings to one another, to connect findings to the possible diagnoses, and to predict additional findings necessary to confirm the diagnosis, even in the absence of prior experience with exactly such a case.⁷⁹

Physicians typically apply hypothetico-deductive approaches when seeing patients with problems outside of their expertise or patients with difficult problems with atypical features within their expertise and apply intuitive pattern recognition for cases within their expertise or less challenging cases. However, diagnostic accuracy appears to depend more on mastery of domain knowledge than on the particular problem-solving method.⁸⁰

Diagnostic Reasoning

There is no correlation between physicians’ ability to collect data thoroughly and their ability to interpret the data accurately.⁸¹ As a decision scientist describes,

You can’t think unless you’ve got distinctions, and we talk about powerful distinctions. I say an expert in anything is an expert because that expert has powerful distinctions. And we don’t mean he or she has memorized the glossary. It’s that you really have a working knowledge of why they’re important and why that distinction was a very valuable one in the history of the subject and in your thinking.⁸²

76. Kassirer et al., *supra* note 70, at 100–04, 255–62; Elstein & Schwartz, *supra* note 70, at 730–31.

77. Trisha Greenhalgh, *Intuition and Evidence—Uneasy Bedfellows?* 52 Brit. J. Gen. Prac. 395, 395 (2002).

78. *Id.* at 395–96.

79. Kassirer et al., *supra* note 70, at 11–12; Elstein & Schwartz, *supra* note 70, at 730–31.

80. Elstein & Schwartz, *supra* note 70, at 730.

81. Arthur S. Elstein & Alan Schwartz, *Clinical Reasoning in Medicine*, in *Clinical Reasoning in the Health Professions* 223, 224 (Joy Higgs et al. eds., 3d ed. 2008).

82. Interview with Ronald Howard on March 9, 2005, The Whitman Institute (document on file with author).

Symptoms and Signs

Although a test is commonly thought of as a sample from a bodily fluid, tissue, or image, a test also could be the presence or absence of a symptom or physical sign. For example, both influenza and the ancestral Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) can present with similar symptoms. However, a distinguishing symptom is the absence of a runny nose, which is nine times more likely with SARS-CoV-2 (95% vs. 11%). Thus, a runny nose occurs in only 5% of patients with SARS-CoV-2 but 89% of patients with influenza, making influenza 18 times more likely in those with a runny nose.⁸³

Bayes' Rule

Over 200 years ago, Reverend Thomas Bayes first wrote a paper, published posthumously, that now forms a critical concept in modern medicine: how to estimate the likelihood of disease following a test result using the likelihood of disease prior to testing and the specific test result obtained. Bayesian analysis refers to a method of combining existing evidence or a pretest probability with additional evidence, for example, from the presence or absence of a positive symptom, sign, test, or research-study result to calculate a posterior probability.⁸⁴

The pretest suspicion of disease or, equivalently, the likelihood or prior probability of disease may be objective, that is, related to incidence (new cases over a specified period of time) or prevalence (existing cases at a particular point in time); based on clinical-prediction rules (e.g., mathematical predictive models to estimate the likelihood of having a stroke or heart attack over the next ten years); or subjective, that is, based on a clinician's subjective estimated likelihood of disease prior to any testing.⁸⁵ Bayes' rule then combines that pretest suspicion with the observed test result. Tests, however, are almost never perfectly accurate. Not everyone with disease has a positive test (i.e., false negatives), and not

83. Nathaniel Hupert et al., *Accuracy of Screening for Inhalational Anthrax After a Bioterrorist Attack*, 139 *Annals Internal Med.* 337, 342 (2003), https://doi.org/10.7326/0003-4819-139-5_part_1-200309020-00009; W. Guan et al., *Clinical Characteristics of Coronavirus Disease 2019 in China*, 382 *NEJM* 1708, 1713 (2020), <https://doi.org/10.1056/NEJMoa2002032>.

84. See David H. Kaye & Hal S. Stern, *Reference Guide on Statistics and Research Methods*, "Bayesian Statistical Methods and Posterior Probabilities," in this manual.

85. See *Gonzalez v. Metro. Transp. Auth.*, 174 F.3d 1016, 1023 (9th Cir. 1999) (describing the implications of Bayes' rule for drug testing and noting that a test with the same false-positive rate will generate a higher proportion of false positives to true positives in a population with fewer drug users); see generally Michael O. Finkelstein & William B. Fairley, *A Bayesian Approach to Identification Evidence*, 83 *Harv. L. Rev.* 489 (1970), <https://doi.org/10.2307/1339656>. For a discussion of Bayesian statistics, see David H. Kaye & Hal S. Stern, *Reference Guide on Statistics and Research Methods*, "Appendix: Conditional Probability and Bayes' Rule," in this manual.

everyone who does not have disease has a negative test (i.e., false positives), which can lead to misestimation of the likelihood of disease.

For example, consider the interpretation of a positive mammogram among women who may have cancer but most do not.⁸⁶ Suppose 1% or 10 of 1,000 40-year-old women have breast cancer. The mammogram will be truly positive in 80% or 8 of the 10 women with breast cancer but falsely positive in 10% or 99 of the 990 women without breast cancer. Among the 107 (8 plus 99) women with a positive mammogram, 8 have breast cancer, so the positive mammogram increases the likelihood of breast cancer from 1% (10 per 1,000) to 7% (8 per 107). The 7% is the probability of breast cancer among 40-year-old women after a positive mammogram, that is, post-test probability of disease after a positive test or predictive-value positive. The above is described as a natural-frequency format for Bayes' rule. When presented as probabilities as in Table 1: There is a 1% probability of breast cancer for these 10,000 women with an 80% probability of a positive mammogram for those with breast cancer and a 10% probability of a positive mammogram without breast cancer. Table 1 shows the calculation of Bayes' rule using probabilities with column C showing the natural frequencies as probabilities and column D multiplying the two probability columns. Column E then divides the column D entries by the total frequency of positive mammograms.

Table 1. A Tabular Form of Bayes' Rule for Test Interpretation

A	B	C	D	E
	Prior Probability	Probability of a Positive Mammogram Given the Condition in Column A	Multiply Columns B and C	Post-test Probability (Column D Divided by Sum)
Breast Cancer	1% = 10/1,000	80% = 8/10	0.8% = 8/1,000	7% = 8/107
No Breast Cancer	99% = 990/1,000	10% = 1/10	9.9% = 99/1,000	93% = 99/107
			Sum = 10.7% = 107/1,000	

Probabilistic Reasoning

There are two types of medical reasoning to infer the likely diagnosis (Figure 2). Traditional medical education teaches disease manifestations by understanding

86. Among 48 physicians, 10% receiving probability data achieved the correct answer versus 46% when given the natural frequency format. Ulrich Hoffrage & Gerd Gigerenzer, *Using Natural Frequencies to Improve Diagnostic Inferences*, 73 Acad. Med. 538 (1998), <https://doi.org/10.1097/00001888-199805000-00024>.

the mechanistic pathways by which the disease causes specific symptoms and signs. For deductive medical inference (deduction), one hypothesizes a specific disease and determines if the manifestations of that disease (e.g., from the literature) account for all of the patient’s observed symptoms and signs.⁸⁷ Deduction, however, assumes the diseases hypothesized contain the most likely set of diagnoses in a specific patient, yet a set of different diseases may have overlapping manifestations and have varying underlying likelihoods in different contexts. In contrast to deduction, inductive medical inference (induction) goes from observing symptoms and signs (what is known) to quantifying the likelihood of each of the possible diagnoses, allowing for a possibly broader consideration of diseases but with uncertainty attached to the likelihood of each disease.

For example, both influenza and SARS-CoV-2 can cause fever, cough, sore throat, nasal congestion, headache, fatigue, muscle and joint aches, chills, nausea, vomiting, and diarrhea. How might Bayes’ rule help distinguish the cause of the symptoms? To illustrate probabilistic inference, the table below presents another tabular form of Bayes’ rule that starts with two diseases (column A) and then the pretest probability of each disease (column B) (e.g., the observed ratio of the incidence of SARS-CoV-2 versus influenza based on prevailing epidemiologic data) followed by the published likelihood of the observing nasal congestion for each disease.⁸⁸ Column D multiplies the numbers in columns B and C to calculate the joint likelihood of those results and the prior probability of each disease. Column E shows that the presence of nasal congestion changes the higher epidemiologic likelihood of SARS-CoV-2 to making influenza much more likely. Bayes’ rule provides the only quantitative approach to incorporate findings to calculate revised likelihoods of diseases given a symptom, sign, or test result through induction (in contrast to the deduction).

Table 2. A Tabular Form of Bayes’ Rule for Diagnosis

A	B	C	D	E
	Prior Probability	Probability of Nasal Congestion Given the Condition in Column A	Multiply Columns B and C	Post-test Probability (Column D Divided by Sum)
SARS-CoV-2	0.75	0.05	0.037	0.14
Influenza	0.25	0.89	0.222	0.86
			Sum = 0.259	

87. This is analogous to assuming the null hypothesis between the intervention and control arm in a randomized trial. One can then deduce the likelihood of the observed results based on that assumption.

88. Guan et al., *supra* note 83, at 1713; Hupert et al., *supra* note 83, at 340.

How well does a symptom distinguish SARS-CoV-2 from influenza? The informative or discriminating ability of a test can be succinctly summarized as a likelihood ratio. The likelihood-ratio positive is the ratio of true-positive rate (sensitivity) to the false-positive rate (one minus the specificity) and expresses how much more likely disease is to be present following a positive test result. Considering influenza as the disease, the likelihood-ratio positive for nasal congestion is 18 ($0.89 \div 0.05$). Likelihood-ratio positives should always exceed one because they increase the likelihood of disease, and likelihood-ratio negatives should always be less than one. Likelihood ratios exceeding 10 or falling below 0.1 are strong discriminators causing “large” changes in the likelihood of disease; those between 5 and 10 or 0.1 and 0.2 cause “moderate” changes; and those between 2 and 5 or 0.2 and 0.5 cause “small” changes.⁸⁹ The likelihood ratios succinctly capture the combined effects of sensitivity and specificity to express the impact of a finding on the likelihood of a disease.

Probabilistic inductive reasoning using Bayes’ rule not only applies to interpreting symptoms and signs for differential diagnosis but also to selection and interpretation of diagnostic tests to exclude or confirm the presence of a disease, thereby avoiding faulty information gathering and processing, for example, faulty interpretation of a test. It also explicitly accounts for varying disease prevalence in different contexts due to patient or population characteristics. For example, a patient with chest pain will have different likelihoods of atherosclerotic disease in the heart if they are being seen in primary care or had been referred by primary care to cardiology outpatient settings. This is because the latter often requires a referral from another doctor who judges the risk of heart disease to be higher, so those patients will have a higher likelihood of worrisome chest pain. This influences the interpretation of a test result as well as the likelihood of disease and the perceived benefit of treatment in the consulting cardiologist.⁹⁰ Beyond differential diagnosis, the prevalence (or pretest probability) of disease in different patient populations affects how well mathematical predictive models perform in other populations (i.e., their generalizability). For example, a logistic regression model that captures the four physical findings that predict strep throat does not perform as well in other populations because the prevalence of strep throat varies by setting.⁹¹

89. David A. Grimes & Kenneth F. Schulz, *Refining Clinical Diagnosis with Likelihood Ratios*, 365 *Lancet* 1500, 1502 (2005), [https://doi.org/10.1016/S0140-6736\(05\)66422-7](https://doi.org/10.1016/S0140-6736(05)66422-7).

90. Harold C. Sox Jr., *The Baseline Electrocardiogram*, 91 *Am. J. Med.* 573 (1991), [https://doi.org/10.1016/0002-9343\(91\)90208-f](https://doi.org/10.1016/0002-9343(91)90208-f).

91. Roy M. Poses et al., *The Importance of Disease Prevalence in Transporting Clinical Prediction Rules: The Case of Streptococcal Pharyngitis*, 105 *Annals Internal Med.* 586, 586 (1986), <https://doi.org/10.7326/0003-4819-105-4-586> (prediction of streptococcal pharyngitis improved when adjusted for the disease prevalence in the specific clinical setting, e.g., by changing the constant in the logistic regression).

Prosecutor's Fallacy

The prosecutor's fallacy involves the misinterpretation of probabilistic information, such as when a prosecutor argues that if the defendant is guilty, then the probability is high that the compelling evidence will be found, and given presentation of that evidence during the trial, the defendant must be guilty. For example, in a highly publicized British case, solicitor Sally Clark was convicted of double homicide of her first two children in part based on testimony from then-eminent pediatrician Professor Sir Roy Meadow, known for Meadow's law that "one cot death [the British term for sudden infant death syndrome or SIDS] is a tragedy, two cot deaths is suspicious and, until the contrary is proved, three cot deaths is murder."⁹² The Confidential Enquiry for Stillbirths and Deaths in Infancy (CESDI) in Britain estimated a SIDS risk at the time of the testimony of 1 in 1,300 live births overall; Sir Meadow applied a lower risk of a SIDS death of 1 in 8,543 births to account for two characteristics of the Clark family that made SIDS death less likely (nonsmoking and affluent parents), but ignoring ecologic bias when accounting for those characteristics and also assuming the same likelihood for both deaths. By squaring 1 in 8,543 births, he testified that the likelihood of two natural SIDS deaths was "one in 73 million" with the potential implication that all other deaths were all by homicide. With 700,000 births annually in the United Kingdom, Sir Meadow testified that a double SIDS death would be expected to occur "once in a hundred years."⁹³

In a commentary, Ray Hill, a professor of mathematics, wrote that some infants die from SIDS and some from homicides. Double SIDS and double infant homicides deaths in the same family are both rare events, so what matters are the "relative chances of these." Using the 700,000 births annually and the incidence data, 538 deaths would be expected from SIDS and 32 from homicide, so on average, SIDS mortality was 17-fold more likely than homicide death. Based on data (albeit sparser data), Hill then estimated the likelihood of a second SIDS death to be 1/228 and a second homicide death to be 1/123, resulting in 2.36 double SIDS deaths and 0.26 double homicide deaths or a 9:1 ratio of double SIDS to double homicide (Table 3).⁹⁴ Hill wondered "whether the *Clark* jury would have

92. Ray Hill, *Multiple Sudden Infant Deaths: Coincidence or Beyond Coincidence?*, 18 Paediatric & Perinatal Epidemiology 320, 320 (2004), <https://doi.org/10.1111/j.1365-3016.2004.00560.x>.

93. *Id.* at 325.

94. Bayes' rule demonstrates how motive in a trial is critical. If the prior probability of an individual committing a crime is zero, despite whatever evidence is presented, the posterior probability of crime by that individual, regardless of the evidence, remains zero. A motive, even if weak, moves the prior probability to be more than zero so that convincing evidence can then raise the likelihood of crime to be more likely than not (i.e., greater than 50%) or, alternatively, beyond a reasonable doubt, e.g., for homicide. For example, if the prior probability of homicide is zero due to the absence of a motive for murder in Table 3, then that zero is multiplied throughout the Homicide row, so the posterior probability for SIDS will always be 1.00 or 100%.

Table 3. Probability of Two Deaths from SIDS vs. Homicides

A	B	C	D	E
	Infant Deaths in 1 Year from Column A	Probability of Second Death	Multiply Columns B–C	Posterior Probability
SIDS	538	1/228	2.36	0.90
Homicide	32	1/123	0.26	0.10
			Sum = 2.62	

convicted if, instead of being given the ‘once in a hundred years figure,’ they had been told that second cot deaths occur about four or five times a year and indeed happen more frequently than second infant murders in the same family.”⁹⁵

Sally Clark was convicted of murder and, after losing her first appeal, won on second appeal in 2003 that would have also allowed appeal based on “flawed statistical evidence” (she died in 2007).⁹⁶ The president of the Royal Statistical Society urged that “statistical evidence be presented only by appropriately qualified statistical experts.”⁹⁷ Nobles and Schiff argued for a simpler approach: “[O]ne does not compensate for the misleading evidence; one simply excludes it.”⁹⁸

The *Clark* case led a Queen’s Counsel to raise the issues of susceptibility toward systematic distortion in forensic science, evident in two forms: “inherent bias towards producing a positive outcome or result” and “the tendency for an adversarial system of investigation and trial to lead to partisan behavior,” particularly “when the experts used in the course of investigation are also the experts used in the trial.”⁹⁹ In the 2003 second appeal of the *Clark* case, the pathologist who changed the cause of death for the first brother after the second death and who also withheld evidence that bacteria were found in the cerebral spinal fluid of the second child stated, “‘It is not my practice to refer to additional results in my post mortem unless they are relevant to the cause of death, as the specimens were referred to another consultant.’ The Court did not consider this approach acceptable. . . . It is in effect the expert trying the case and deciding what is relevant evidence, not the court.”¹⁰⁰

95. Hill, *supra* note 92, at 325.

96. Clare Dyer, *Falsely Convicted Sally Clark Dies Suddenly*, 334 B.M.J. 602, 604 (Mar. 24, 2007), <https://doi.org/10.1136/bmj.39160.770637.DB>.

97. Richard Nobles & David Schiff, *Misleading Statistics Within Criminal Trials*, 47 Med. Sci. Law 7, 10 (2007), <https://doi.org/10.1258/rsmmsl.47.1.7>.

98. *Id.*

99. Clare Montgomery, *Forensic Science in the Trial of Sally Clark*, 44 Med. Sci. L. 185, 185 (2004), <https://doi.org/10.1258/rsmmsl.44.3.185>.

100. A.R.W. Forrest, *Sally Clark: A Lesson for Us All*, 43 Sci. Just. 63, 64 (2003), [https://doi.org/10.1016/S1355-0306\(03\)71744-4](https://doi.org/10.1016/S1355-0306(03)71744-4).

Multiple Diseases and Multiple Tests

The terms sensitivity and specificity apply to a simple situation in which disease is present or absent, and a test can be positive or negative, but terminology and interpretation become more complicated when multiple diseases are under consideration and when multiple test results may occur.¹⁰¹ For example, consider blood in the urine (hematuria), which could be caused by a urinary tract infection, a kidney stone, or a bladder cancer, among many other diseases. The terms sensitivity and specificity are no longer appropriate because disease is not simply present or absent. Instead, they are replaced by the term conditional probabilities; that is, sensitivity is replaced by the likelihood of blood in the urine with a urinary tract infection, or with a kidney stone, or with a bladder cancer. Similarly, a very positive test has a different interpretation than a weakly positive test, and Bayes' rule can quantify the difference. Results from multiple tests can be combined with Bayes' rule by applying Bayes' rule to the first test result and then reapplying Bayes' rule to subsequent test results with the post-test probability after the first test becoming the pretest for the second test. This approach assumes that the result of the first test does not affect the test characteristics (sensitivity or specificity) of the second test (i.e., that there is conditional independence of each test). When two tests are available, the screening will usually occur first with the high-sensitivity test to detect a high proportion of those with disease (true positives, or "ruling in" disease). Those with a positive first test will then undergo a high-specificity test to reduce the number of individuals who do not have disease ("ruling out" disease) in those with a false-positive first test.

The choice of whether to do a test will depend on the pretest probability. The basic rule of thumb is to do a test only if it could change the action you would take. With a high pretest probability, you would treat unless a test result would drive the probability to such a low level that you wouldn't treat, which would require a very sensitive test (few false negatives). If the pretest probability is low, you wouldn't treat unless the post-test probability was high, which requires a very specific test (few false positives).

Genomic Analyses

Bayes' rule becomes even more relevant in the genomic-medicine era.¹⁰² Suppose a genetic test has a sensitivity and specificity of 99.9%, and suppose the probability of disease is 1 in 1,000 if a positive family history is present and 1 in 100,000 if no

101. Kassirer et al., *supra* note 70, at 21–22. See also Steve C. Gold et al., *Reference Guide on Epidemiology*, "Specificity of the Association," in this manual.

102. Isaac S. Kohane et al., *The Incidentalome: A Threat to Genomic Medicine*, 296 JAMA 212, 213 (2006), <https://doi.org/10.1001/jama.296.2.212>.

family history is present. Screening 1,000 individuals with a positive family history for the gene results in two positive tests: one individual truly has a disease, and in the other, the test is a false positive. Screening 10 million individuals without a family history results in 10,100 positive tests, of which 100 individuals have a disease and 10,000 do not. Even with a specificity of 99.99%, if a test screens for 10,000 genes simultaneously, then 63% of individuals will have at least one false-positive test result. Based simply on the genetic test results alone, neither individuals nor physicians would be able to distinguish those with true-positive results from those with false-positive results, thereby potentially leading to inappropriate monitoring or treatment for all with positive test results.

Positivity Criterion

The sensitivity and specificity of a quantitative test rely on setting a positivity criterion, a level for determining what is negative or positive for disease. When the criterion is made stricter (a higher level) for a test where higher is more positive, then sensitivity falls and specificity increases, and if the criterion is made laxer (a lower level), then sensitivity rises, and specificity falls. Depending on the context of the testing, it may be more appropriate to choose a laxer criterion (e.g., screening donated blood for HIV infection where the benefit is reducing transfusion-associated HIV transmission, and the harm is discarding some uninfected units of donated blood) or a stricter one (e.g., screening a low-prevalence population for HIV infection where the benefit is reducing false-positive diagnoses from higher specificity, and the harm is missing some truly HIV-infected individuals from lower sensitivity).¹⁰³ Thus the benefits of finding and treating a person with disease versus the risk of harm from mislabeling or treating a person without disease should help establish what should be considered negative or positive.

Bayesian vs. Frequentist Statistics

In a randomized trial, frequentist statistics involve assuming a null hypothesis (no difference between an intervention and control) and then calculating the likelihood of the outcomes observed. For example, using a fair coin (null hypothesis), the likelihood of two heads in a row occurring would be 25% (HT, HH, TH, TT), of three heads in a row 12.5%, of four heads in a row 6.25%, and of five heads in a row 3.125%. The alpha or *p*-value for statistical significance is typically 5%, and some argue that four or five heads in a row would make some suspicious of the coin being unfair. The American Statistical Association has, for

103. Klemens M. Meyer & Stephen G. Pauker, *Screening for HIV: Can We Afford the False Positive Rate?*, 317 NEJM 238, 240 (1987), <https://doi.org/10.1056/NEJM198707233170410>.

the first time, made a statement about statistical practice methods to move away from $p < 0.05$:

Don't believe that an association or effect exists just because it was statistically significant. Don't believe that an association or effect is absent just because it was not statistically significant. Don't believe that your p -value gives the probability that chance alone produced the observed association or effect or the probability that your test hypothesis is true.¹⁰⁴

One of the most highly cited papers in medicine, titled *Why Most Published Research Findings Are False*, states, "It can be proven that most claimed research findings are false. . . . a consequence of the . . . ill-founded strategy of claiming conclusive research solely on the basis of a single study assessed by formal statistical significance, typically for a p -value less than 0.05."¹⁰⁵ When a study finds a surprising result with $p = 0.05$, the overwhelming majority of physicians "will confidently state there is a 95% or greater chance that the null hypothesis is incorrect . . . an understandable but categorically wrong interpretation because the P value" assumes "the null hypothesis is true. It cannot, therefore, be a direct measure of the probability that the null hypothesis is false."¹⁰⁶

Bayesian inference is often described as showing "how belief is altered by data" or "belief calculus" and being subjective, especially when no prior evidence exists, whereas others term it "evidential calculus" by combining the prior odds of the null hypothesis with a Bayes factor as the "weight of the evidence,"¹⁰⁷ for example, from a randomized trial.

The Bayes factor differs in many ways from a P value. First, the Bayes factor is not a probability itself but a likelihood ratio of probabilities, i.e., Probability of the Data, given the null hypothesis divided by the Probability of the Data, given the alternative hypothesis, so it can vary from zero to infinity. It requires two hypotheses, making it clear that for evidence to be against the null hypothesis, it must be for some alternative. Second, the Bayes factor depends on the probability of the observed data alone . . . so we begin to understand what represents strong evidence, and weak evidence . . .¹⁰⁸

Standard frequentist statistical and Bayesian methods have tradeoffs, with standard frequentist methods weakest when drawing conclusions from a single

104. Ronald L. Wasserstein et al., *Moving to a World Beyond " $p < 0.05$,"* 73 Am. Statistician 1, 1 (2019), <https://doi.org/10.1080/00031305.2019.1583913>.

105. John P. A. Ioannidis, *Why Most Published Research Findings Are False*, 2 PLoS Med. e124, 0696 (2005), <https://doi.org/10.1371/journal.pmed.0020124>.

106. Goodman, *supra* note 71, at 998.

107. Steven N. Goodman, *Toward Evidence-Based Medical Statistics. 2: The Bayes Factor*, 130 Annals Internal Med. 1005, 1005 (1999), <https://doi.org/10.7326/0003-4819-130-12-199906150-00019>; see generally Liesa L. Richter & Daniel J. Capra, *The Admissibility of Expert Testimony*, "Weight of the Evidence Analysis: Closing the Analytical Gap," in this manual.

108. Goodman, *supra* note 107, at 1006.

experiment, whereas Bayesian methods make inductive inference from that experiment. Bayesian methods, however, use prior probability distributions and are imperfect subjective estimates of what is known or unknown, so that over multiple experiments, Bayesians cannot guarantee that their 95% credible intervals (the Bayesian analog of the 95% confidence interval) will hold up. Thus, frequentist criteria can and have been used to evaluate Bayesian and likelihood methods. “Putting *P* values aside, Bayesian and frequentist approaches each provide an essential perspective that the other lacks. The way in which we balance their sometimes conflicting demands is what makes the process of learning from nature creative, exciting, uncertain, and, most of all, human.”¹⁰⁹

Causal Reasoning

While considering risk factors (e.g., age, gender, exposures such as smoking, family history, or medical conditions), physicians will often use any type of evidence¹¹⁰ that might support causation—for example, biological plausibility,¹¹¹ physiological drug effects, case reports, or temporal proximity¹¹² to an exposure.¹¹³ Although physicians use epidemiological studies in their decision-making, “they are accustomed to using *any* reliable data to assess causality, no matter what their source” because they must make care decisions (e.g., treat or not) even in the face of uncertainty.¹¹⁴ This is in contrast to the courts, which require a higher standard than clinicians or regulators; causation cannot just be “possible,” rather “a ‘preponderance of evidence’ establishes that an injury was caused by an alleged exposure.”¹¹⁵ For physicians, causal reasoning typically involves understanding how abnormalities in physiology, anatomy, genetics, or biochemistry lead to the clinical manifestations of disease. Through such reasoning, physicians develop a causal cascade or “chain or web of causation” linking a sequence of plausible cause-and-effect mechanisms to arrive at the pathogenesis or pathophysiology of a disease. For example, kidney failure leads to poor drug

109. *Id.* at 1012.

110. Jerome P. Kassirer & Joe S. Cecil, *Inconsistency in Evidentiary Standards for Medical Testimony: Disorder in the Courts*, 288 JAMA 1382, 1384 (2002), <https://doi.org/10.1001/jama.288.11.1382>; see also section titled “Evidence-Based Medicine” below for levels of evidence.

111. See *Keenan v. Sec’y of Health & Hum. Servs.*, No. 99–561V, 2007 WL 1231592 (Fed. Cl. Apr. 5, 2007).

112. *But see Wilson v. Taser Int’l, Inc.*, 303 F. App’x 708, 714 (11th Cir. 2008) (“[A]lthough a doctor usually may primarily base his opinion as to the cause of a plaintiff’s injuries on this history where the patient ‘has sustained a common injury in a way that it commonly occurs,’ . . . Dr. Meier could not rely upon the temporal connection between the two events to support his causation opinion in this case.”).

113. Kassirer & Cecil, *supra* note 110, at 1384.

114. *Id.*

115. *Id.*

excretion, resulting in symptoms or signs of drug toxicity.¹¹⁶ These causal-reasoning pathways center on disease manifestations to help treating clinicians, who are commonly less concerned with distinguishing causation from association, with diagnosis and treatment (e.g., Sir Austin Bradford Hill's considerations for inferring causation).¹¹⁷

Pattern recognition of concomitant symptoms and signs can trigger a diagnosis. For example, cough, lung lesions, and enlarged breasts (gynecomastia) in a thirty-seven-year-old man could spark the diagnosis of metastatic germ cell cancer.¹¹⁸ More typically, physicians use causal reasoning in diagnostic refinement and verification to examine a diagnosis for its coherency, namely, asking whether its physiological mechanism would be expected to lead to the observed manifestations and whether it adequately accounts for all normal and abnormal findings and the disease time course. Once treatment has been implemented, clinicians must make causal judgments to determine whether alteration in patient status is the result of progression of disease or an adverse consequence of treatment, or whether the absence of improvement results should prompt a change in therapy or even reconsideration of the diagnosis.

Pathophysiological reasoning, however, also can lead to incorrect conclusions. In patients with heart failure with a weakened heart, a class of medications called beta blockers had been thought to be contraindicated because beta blockers would decrease the strength of the heart muscle contraction and slow the then thought to be compensatory increase in heart rate. Subsequent studies found that beta blockers in patients with heart failure actually increased survival without ill effect. Similarly, physicians once thought that atherosclerotic blockages in heart arteries slowly progressed to cause a heart attack, so that revascularizing those plaques through heart bypass surgery would prevent heart attacks.¹¹⁹ Over the past twenty-five years, however, scientific evidence has emerged that small vulnerable atherosclerotic plaques (not amenable to revascularization because of their small size) can suddenly rupture and cause heart attacks. Not surprisingly, revascularization trials involving either bypass surgery or percutaneous interventions such as stenting or angioplasty do relieve symptoms but do not diminish the risk of having a heart attack or improve survival for most patients.¹²⁰

116. *Id.*

117. See generally *Reference Guide on Exposure Science and Exposure Assessment*, *Reference Guide on Epidemiology*, *Reference Guide on Toxicology*, and the prologue to those three guides, in this manual.

118. Kassirer et al., *supra* note 70, at 29.

119. David S. Jones, *Visions of a Cure: Visualization, Clinical Trials, and Controversies in Cardiac Therapeutics, 1968–1998*, 91 *Isis* 504, 532 (2000), <https://doi.org/10.1086/384853>.

120. Thomas A. Trikalinos et al., *Percutaneous Coronary Interventions for Non-Acute Coronary Artery Disease: A Quantitative 20-Year Synopsis and a Network Meta-Analysis*, 373 *Lancet* 911, 915 (2009), [https://doi.org/10.1016/S0140-6736\(09\)60319-6](https://doi.org/10.1016/S0140-6736(09)60319-6); Mark A. Hlatky et al., *Coronary Artery Bypass Surgery Compared with Percutaneous Coronary Interventions for Multivessel Disease: A Collaborative*

Although treating physicians¹²¹ may testify on both general and specific causation, as with use of evidence for causation, their standards for evidence vary.¹²² For example, some physicians may stop using a drug after the first reports of adverse effects, and others may continue to use a drug despite evidence of harm from randomized controlled trials. Determining whether an effect is a class effect or is drug-specific can be difficult. In a randomized trial limited to patients with documented weakened hearts, one particular beta blocker was found not to confer a survival benefit, and as a result, the heart failure guidelines limited their recommendation to three beta-blocker drugs with documented mortality benefit in trials.¹²³

Treating physicians may be aware of patient-specific risk factors such as smoking or family history but may not routinely review specialized aspects of such data, for example, toxicology, industrial hygiene, environment, and some aspects of epidemiology. Additional experts may assist in distinguishing general from specific causation by using specialized knowledge to weigh the relative contribution of each putative causative factor to determine “reasonable medical certainty” or “reasonable medical probability.” The determination of general causation involves medical and scientific literature review and the evaluation of epidemiological data, toxicological data, and dose–response relationships.

Causal Inference Methods

It is important to recognize that

[S]cientists report that an evaluation of data and scientific evidence to determine whether an inference of causation is appropriate requires judgment and interpretation. Scientists are subject to their own value judgments and preexisting biases . . . although one scientist or group of scientists comes to one conclusion about factual causation, they recognize that another group that comes to a contrary conclusion might still be “reasonable.” Courts, thus, should be cautious about adopting specific “scientific” principles, taken out of context, to

Analysis of Individual Patient Data from Ten Randomised Trials, 373 *Lancet* 1190, 1192 (2009), [https://doi.org/10.1016/S0140-6736\(09\)60552-3](https://doi.org/10.1016/S0140-6736(09)60552-3).

121. See generally *Bland v. Verizon Wireless, LLC*, 538 F.3d 893 (8th Cir. 2008) (upholding district court’s decision to reject a treating physician’s evidence of causation under *Daubert*); *Cripe v. Henkel Corp.*, 318 F.R.D. 356 (N.D. Ind.), *aff’d*, 858 F.3d 1110 (7th Cir. 2017) (indicating that a treating doctor is providing expert testimony if doctor offers opinions based on scientific or technical knowledge).

122. Kassirer & Cecil, *supra* note 110, at 1384.

123. Paul A. Heidenreich et al., *2022 AHA/ACC/HFSA Guideline for the Management of Heart Failure: A Report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines*, 145 *Circulation* e895, e931 (2022), <https://doi.org/10.1161/CIR.0000000000001062>.

formulate bright-line legal rules or conclude that reasonable minds cannot differ about factual causation.¹²⁴

In an emerging area, “[b]ecause identifying individual causal effects is generally not possible, we now turn our attention to an aggregated causal effect: the average causal effect in a population of individuals.”¹²⁵ Consider, for example, hormone replacement therapy for postmenopausal women. Multiple observational studies using methods such as case-control, cross-sectional, and cohort designs¹²⁶ suggested an association between hormone therapy and reduction in heart attack, but such designs are subject to confounding and bias and are particularly weak for causation because in case-control and cross-sectional studies, the sequence of the exposure and outcome is unknown.

To resolve the question, the Women’s Health Initiative (WHI) study randomized women to hormone replacement therapy or placebo and found a statistically significant increase in harmful clot-related disorders—heart attack, stroke, and heart-related mortality—over five years, but evident in the first year after initiation of hormone therapy.¹²⁷ Heart attacks are caused by blood clots and plaque rupture, and so the results were consistent with the known biological mechanism of estrogens in the clotting cascade. However, patients in the WHI study were, on average, sixty-three years old and therefore not perimenopausal, as were the participants analyzed in the observational studies. In a novel approach, researchers emulated the design and intention-to-treat (ITT) analysis aspect of the WHI randomized trial in the prospective observational Nurses’ Health Study and found that hormone-replacement treatment effects were similar to those from the randomized trial, suggesting that “the discrepancies between WHI and the Nurses’ Health Study ITT estimates could be largely explained by differences in the distribution of time since menopause and length of follow-up.”¹²⁸ These emerging methods for analyzing non-randomized study designs address “[t]he fundamental problem of causal inference . . . that at most one of the potential outcomes can be realized and thus observed” as a consequence of an action (e.g., taking a pill).¹²⁹

124. See Restatement (Third) of Torts, § 28 cmt. c, at 403 (2010).

125. Miguel A. Hernán & James M. Robins, *Causal Inference: What If* 4 (2020).

126. See Steve C. Gold et al., *Reference Guide on Epidemiology*, “Types of Observational Study Design,” in this manual.

127. Jacques E. Rossouw et al., *Risks and Benefits of Estrogen Plus Progestin in Healthy Postmenopausal Women: Principal Results from the Women’s Health Initiative Randomized Controlled Trial*, 288 JAMA 321, 327–29 (2002), <https://doi.org/10.1001/jama.288.3.321>; JoAnn E. Manson et al., *Estrogen Plus Progestin and the Risk of Coronary Heart Disease*, 349 NEJM 523, 527–28 (2003), <https://doi.org/10.1056/NEJMoa030808>.

128. Miguel A. Hernán et al., *Observational Studies Analyzed Like Randomized Experiments: An Application to Postmenopausal Hormone Therapy and Coronary Heart Disease*, 19 Epidemiology 766, 766 (2008), <https://doi.org/10.1097/EDE.0b013e3181875e61>.

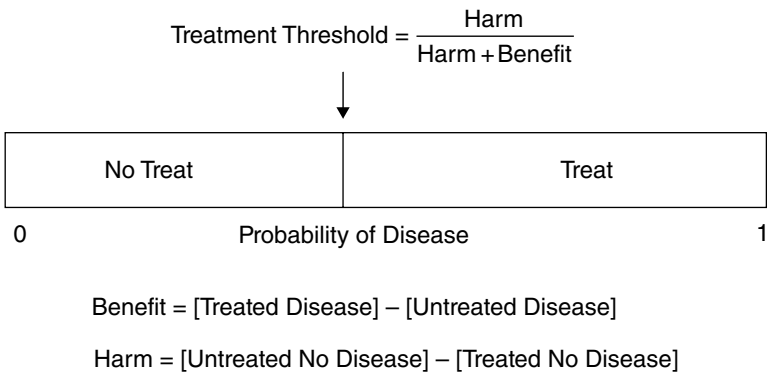
129. Guido W. Imbens & Donald B. Rubin, *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction* 6 (2015), <https://doi.org/10.1017/CBO9781139025751>.

Treatment Decision-Making

Treatment Thresholds

Sir William Osler describes medicine as “a science of uncertainty and an art of probability.”¹³⁰ Through diagnostic reasoning, clinicians seek to identify the cause of a patient’s complaints and findings to determine the most appropriate therapy to alleviate symptoms or prolong life—including no therapy, as ineffective therapy carries risk for harm. In 1975, Pauker and Kassirer described the threshold approach to treatment decisions.¹³¹ For a likelihood of a disease ranging from zero (absent) to one (certainty), they determined that the threshold probability of disease above which patients should be treated depended on the benefits and harms of treating patients with and without disease (Figure 4). The benefit (B) equaled the difference in outcomes of treatment in patients with disease minus their outcomes without treatment; the harm (H) equaled the outcomes in patients without disease who are not treated minus their outcomes if treated. The threshold probability equals H divided by (H + B). If the benefit is high, then the threshold for treatment is low (e.g., antibiotics for pneumonia); if the harm is high, greater diagnostic certainty must exist before treatment (e.g., chemotherapy for cancer usually requires tissue diagnosis). For the diagnostic dilemma of a sarcoidosis or tuberculosis case, tuberculosis (TB) treatment would confer a 30% survival benefit with a 0.15% risk of mortality from TB treatment, so the threshold for TB treatment equals $0.15\% / (0.15\% + 30\%) = 0.15\% / 30.15\% = 0.5\%$, far below the 70% estimated likelihood prior to testing by the discussant.¹³²

Figure 4. Treatment threshold.



130. Mark E. Silverman et al., *The Quotable Osler* 46 (2003).

131. Stephen G. Pauker & Jerome P. Kassirer, *Therapeutic Decision Making: A Cost-Benefit Analysis*, 293 NEJM 229, 230–31 (1975), <https://doi.org/10.1056/NEJM197507312930505>.

132. Richard I. Kopelman et al., *Clinical Problem-Solving: A Little Math Helps the Medicine Go Down*, 341 NEJM 435, 437–38 (1999), <https://doi.org/10.1056/NEJM199908053410608>.

Testing Decision-Making

Testing Thresholds

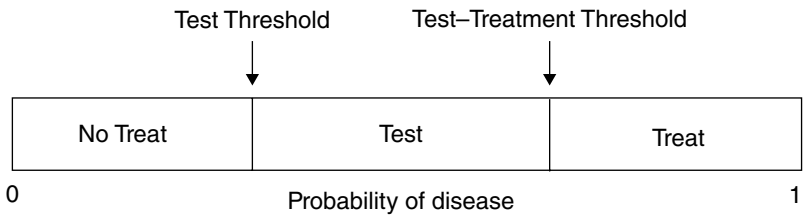
When should a test be done? A guiding principle is that a test should be performed only if the test result could change the care plan. Thus, if a positive test moves the post-test probability of disease across the treatment threshold from a pretest probability falling in the no-treat range, it should change the plan from no treat to treat (or conversely with a negative test when the pretest probability falls in the treat range). The no-treat-test threshold is the lowest pretest probability at which the post-test probability equals the treatment threshold. When a post-test probability does not cross an action threshold, generally it should not be done.

In 1980, Pauker and Kassirer extended their threshold approach to treatment decisions by defining conceptual thresholds for testing (Figure 5).¹³³ They determined that the threshold probabilities at which patients should be tested or treated (or neither) depended on treatment benefit (B) and harm (H) in patients with and without disease but also the sensitivity and specificity of the test as calculated by the likelihood-ratio positive ($LR+ = \text{sensitivity} / (1 - \text{specificity})$) and likelihood-ratio negative ($LR- = (1 - \text{sensitivity}) / \text{specificity}$). The test threshold above which testing should be done equals $H / (H + (LR+) \times B)$. Because $LR+$ increases the likelihood of disease, it always exceeds one, so it magnifies the benefit of treatment and makes the test threshold lower than the treat threshold. Similarly, the $LR-$ is always less than one, so it reduces benefit, thereby raising the test-treatment threshold above the treatment threshold.¹³⁴ In the diagnostic dilemma of TB or sarcoidosis, the $LR-$ for a negative TB skin test (for TB) is 0.26 (0.25/0.95), yielding a test-treatment threshold of 1.9%, and the $LR-$ for an abnormal angiotensin-converting enzyme (ACE) blood test is 0.06 (0.05/0.8), yielding a test-treatment threshold of 7.4%, and even with both $LR-$ tests, the test-treatment threshold rises only to 23% with the discussant's 70% estimate exceeding all three of these thresholds, supporting empiric TB treatment. In summary, the treatment-threshold probability and the $LR-$ and $LR+$ of the test divide the probability scale into three parts: no treat, test, and treatment. If the pretest probability is known, the preferred next step can be identified.

133. Stephen G. Pauker & Jerome P. Kassirer, *The Threshold Approach to Clinical Decision Making*, 302 NEJM 1109, 1110–11 (1980), <https://doi.org/10.1056/NEJM198005153022003>.

134. Kopelman et al., *supra* note 132, at 437.

Figure 5. Test and test–treatment thresholds.



$$\text{Test Threshold} = \frac{\text{Harm}}{\text{Harm} + (\text{LR}+) \times \text{Benefit}}$$

$$\text{Test–Treatment Threshold} = \frac{\text{Harm}}{\text{Harm} + (\text{LR}–) \times \text{Benefit}}$$

Screening

As articulated by Wilson and Jungner for the World Health Organization in 1968,

[t]he object of screening for disease is to discover those among the apparently well who are in fact suffering from disease. They can then be placed under treatment and, if the disease is communicable, steps can be taken to prevent them from being a danger to their neighbours. In theory, therefore, screening is an admirable method of combating disease, since it should help detect it in its early stages and enable it to be treated adequately before it obtains a firm hold on the community.¹³⁵

But as they also note, “[i]n practice, there are snags.”¹³⁶

Decades ago, routine screening tests included ordering anywhere from six to twelve blood tests as part of an annual visit in asymptomatic patients. Normal ranges for biochemical tests are often based on the 95% confidence intervals in a healthy population. By convention, values outside the 2.5% lower and upper extremes are considered to be abnormal, so ordering six blood tests in a healthy individual yields a 74% chance that all six tests will be normal (0.95⁶), that is, a 26% chance that one or more tests may be abnormal, assuming each test is independent of the others. Similarly, when ordering twelve tests in a normal person, there is a 54% chance that all twelve will be normal, so a 46% chance that one or more will be abnormal. Simply ordering tests in healthy individuals in the absence of clinical suspicion of a disease may result in many false-positive test

135. J.M.G. Wilson & G. Jungner, Principles and Practice of Screening for Disease No. 34, at 7, https://iris.who.int/bitstream/handle/10665/37650/WHO_PHP_34.pdf.

136. *Id.*

results that can lead to false alarms, anxiety, additional testing, and possible morbidity or mortality from subsequent testing or interventions.¹³⁷

Even a valueless screening test may appear to be beneficial because of lead-time bias. That is, if screened and unscreened patients have the same prognosis from the time of onset of symptoms to death, screened patients only appear to live longer because the time elapsed from diagnosis by screening to death exceeds the time elapsed from time of symptom onset to death.¹³⁸ A second bias, length bias, also leads to overestimating the benefit from screening.¹³⁹ A randomized trial of screening or no screening is conducted over a limited duration from study initiation to termination. Among the unscreened patients, disease only becomes evident through the development of symptoms, which would be more likely with the aggressive form of the disease, conferring a poorer prognosis for these patients. Among the screened patients, the screening test detects patients with both aggressive and indolent forms of the disease. Some of the patients with indolent cancer may not develop symptoms until after the trial ends. These slower growing cases can make the patients detected by screening appear to live longer, enriching the prognosis of the screening group with those patients with indolent cancers.

Screening can also result in pseudo-disease or overdiagnosis, such as the identification of slow-growing cancers that, even if untreated, would never cause symptoms or reduce survival.¹⁴⁰ Although lung cancer is commonly thought to be one of the more aggressive cancers, an autopsy study found that one-third of lung cancers were unsuspected prior to autopsy, and nearly all of these patients with unsuspected lung cancer prior to autopsy died from other causes.¹⁴¹ The lifetime risk of dying from prostate cancer is about 3%, yet 60% of men in their sixties have prostate cancer, and so, screening and detecting all men with prostate cancer in their sixties would lead to the treatment of many men who would not have died from prostate cancer.¹⁴²

137. A radiologist described his own experience to illustrate the clinical aphorism that “the only ‘normal’ patient is one who has not yet undergone a complete work-up.” He had a negative CT scan of the colon, but the CT scan also provided images outside the colon with radiologists identifying lesions in the kidneys, liver, and lungs. This resulted in additional CT scans, a liver biopsy, PET scan, video-aided thoracoscopy (a flexible scope inserted into the chest), and three wedge resections of the lung leading to multiple tubes, medications, and “excruciating pain” that required five weeks for recovery. William J. Casarella, *A Patient’s Viewpoint on a Current Controversy*, 224 *Radiology* 927 (2002), <https://doi.org/10.1148/radiol.2243020024>.

138. Black, *supra* note 66, at 1280.

139. *Id.*

140. *Id.* at 1280–81.

141. Charles K. Chan et al., *More Lung Cancer but Better Survival: Implications of Secular Trends in “Necropsy Surprise” Rates*, 96 *Chest* 291, 293 (1989), <https://doi.org/10.1378/chest.96.2.291>.

142. Karsten J. Jorgensen & Peter C. Gotzsche, *Overdiagnosis in Publicly Organised Mammography Screening Programmes: Systematic Review of Incidence Trends*, 339 *B.M.J.* b2587 at 1 (2009), <https://doi.org/10.1136/bmj.b2587>; Michael J. Barry, *Prostate-Specific-Antigen Testing for Early Diagnosis of Prostate Cancer*, 344 *NEJM* 1373, 1373 (2001), <https://doi.org/10.1056/NEJM200105033441806>.

Diagnostic Testing

Based on patients' history and physical examination, clinicians establish diagnostic possibilities. They may then request additional tests to reduce uncertainty and to confirm the diagnosis as part of diagnostic verification. Although theoretically all tests could be ordered, tests should be chosen based on a clinical suspicion because of possible morbidity or even mortality from inappropriate testing. Normative, prescriptive decision models for reasoning in the presence of uncertainty suggest that whether and which tests get ordered should depend on test sensitivity and specificity and the benefit and harm of treatment as discussed in the section titled "Diagnostic Reasoning" above; when testing carries risks for morbidity or mortality harms, the probabilities of disease for which testing should be done become narrower, and so physicians should be more likely to treat empirically or neither test nor treat.¹⁴³ As sensitivity and specificity increase, the range of probabilities in which testing should be done expands.

Although an abnormal test result may be found, that abnormality may not be causing symptoms. For example, herniated lower back (lumbar spine) discs are found in approximately 25% of healthy individuals without back pain as an incidental finding. If signs such as a foot drop develop, additional muscle and nerve conduction studies could confirm evidence of nerve compromise from the herniated disc, but such tests are painful. Over time, sequential images show that herniated discs have a partial or complete resolution after six months without surgery as the disc shrinks. Therefore, a herniated disc may be seen with CT or MRI scanning in patients with or without symptoms, so just having symptoms and evidence of a herniated disc would not be a sufficient indication for back surgery.¹⁴⁴ In the absence of severe or progressive neurological deficits, elective disc surgery could be considered for patients with probable herniated discs who have persistent symptoms and findings consistent with sciatica (not just low back pain) for four to six weeks, but such "patients should be involved in decision making" (see section titled "Patient care" below).¹⁴⁵

Tests initially felt to be useful may be found to be less valuable over time.¹⁴⁶ Among other potential biases,¹⁴⁷ this may occur because of the choice of the

143. Stephen G. Pauker & Jerome P. Kassirer, *The Threshold Approach to Clinical Decision Making*, 302 NEJM 1109, 1110, 1115–16 (1980), <https://doi.org/10.1056/NEJM198005153022003>.

144. Richard A. Deyo & James N. Weinstein, *Low Back Pain*, 344 NEJM 363, 366–67 (2001), <https://doi.org/10.1056/NEJM200102013440508>.

145. *Id.* at 368.

146. David F. Ransohoff & Alvan R. Feinstein, *Problems of Spectrum and Bias in Evaluating the Efficacy of Diagnostic Tests*, 299 NEJM 926, 926 (1978), <https://doi.org/10.1056/nejm197810262991705>.

147. Penny Whiting et al., *Sources of Variation and Bias in Studies of Diagnostic Accuracy: A Systematic Review*, 140 Annals Internal Med. 189, 192–94 (2004), <https://doi.org/10.7326/0003-4819-140-3-2004020430-00010>.

study population used to determine the test's sensitivity and specificity. For example, an FDA-approved rapid test for HIV infection has a reported specificity of 100%, implying that any positive tests must indicate truly infected individuals, yet one of the populations in which testing is recommended is women who have had prior children and are in labor but have not yet had an HIV test during the pregnancy.¹⁴⁸ In fifteen multiparous women, this rapid HIV test resulted in one false-positive test result, yielding a specificity of 93%,¹⁴⁹ and so not all pregnant women with positive rapid HIV tests can be assumed to be truly infected.

Biomarker or Image Testing

Once a diagnosis has been established, additional biomarker testing may be performed to establish the extent of disease (e.g., staging of cancer), monitor response to therapy, or direct treatment. In women with breast cancer, for example, finding a genetic marker called the human epidermal growth factor receptor type 2 (HER2, also called HER2/neu) gene identified patients who responded poorly to any of the standard chemotherapeutic agents resulting in a poor prognosis. Illustrative of the era of pharmacogenomics, adjuvant chemotherapy combined with a monoclonal antibody in HER2-positive breast cancer patients has been found to delay progression and prolong survival.¹⁵⁰ For the treatment of lymphomas, monitoring of disease response can be done through imaging. The lack of response has been used to indicate treatment failure and to recommend a change in chemotherapy.

Variation and Standards in Medicine

Variation in Medical Care

Studies over the past several decades show substantial geographic variation in the utilization rates for medical care within small areas or local regions (e.g., a twelve-fold variation in the use of tonsillectomy in rural Vermont) and between large areas or widespread regions (e.g., a four- to five-fold variation in the performance of other discretionary surgical procedures such as hip replacement, coronary

148. Food and Drug Administration, OraQuick Rapid HIV-1 Antibody Test, <https://www.fda.gov/media/73699/download>.

149. *Id.* at 14.

150. Dennis J. Slamon et al., *Use of Chemotherapy Plus a Monoclonal Antibody Against HER2 for Metastatic Breast Cancer That Overexpresses HER2*, 344 NEJM 783, 787–89 (2001), <https://doi.org/10.1056/NEJM200103153441101>.

bypass surgery, and prostatectomy, among others).¹⁵¹ Even when limiting the analysis to seventy-seven U.S. hospitals with reputations for high-quality care in managing chronic illness, the care that patients received in their last six months of life varied extensively, ranging from hospital stays of nine to twenty-seven days (three-fold variation), intensive-care unit stays of two to ten days (five-fold variation); and physician visits of eighteen to seventy-six (four-fold variation), depending on the hospital at which patients received their care.¹⁵²

Four categories of variation are recognized: (1) underuse of effective care, (2) issues of patient safety, (3) concern for preference-sensitive care, and (4) notions of supply-sensitive services.¹⁵³ Effective care refers to treatments that are known to be beneficial and that nearly all patients should receive with little influence of patient preferences—for example, the use of beta blockers following myocardial infarction. The underuse of effective care was illustrated by one prominent study that identified 439 high-quality process measures for 30 conditions and preventive care. In assessing the use of measures that were clearly recommended (i.e., clearly beneficial), they found that only about 50% of patients received these highly recommended care processes.¹⁵⁴ Issues of patient safety refer to the execution of care and the occurrence of iatrogenic complications (i.e., complications resulting from healthcare interventions). The IOM emphasized that “[g]etting the right diagnosis is a key aspect of healthcare: it provides an explanation of a patient’s health problem and informs healthcare decisions,” yet “[i]t is likely that most of us will experience at least one diagnostic error in our lifetime, sometimes with devastating consequences.”¹⁵⁵ Concern for preference-sensitive care refers to treatment choices that should depend on patient health goals or preferences.¹⁵⁶ Prostate surgery helps relieve symptoms of an enlarged prostate (such as frequent urination and waking up at night to urinate) but carries a risk of losing sexual function. Separate from the probability of losing sexual function, in preference-sensitive care, the decision to have prostate surgery depends on how much the enlarged prostate symptoms bother the patient and on how important sexual function is to them, that is, it depends on their preferences and values.¹⁵⁷ Finally,

151. John D. Birkmeyer et al., *Understanding Regional Variation in the Use of Surgery*, 382 *Lancet* 1121, 1122 (2013), [https://doi.org/10.1016/S0140-6736\(13\)61215-5](https://doi.org/10.1016/S0140-6736(13)61215-5).

152. John E. Wennberg et al., *Use of Hospitals, Physician Visits, and Hospice Care During Last Six Months of Life Among Cohorts Loyal to Highly Respected Hospitals in the United States*, 328 *B.M.J.* 607, 607 (2004), <https://doi.org/10.1136/bmj.328.7440.607>.

153. John E. Wennberg, *Unwarranted Variations in Healthcare Delivery: Implications for Academic Medical Centres*, 325 *B.M.J.* 961, 962–63 (2002), <https://doi.org/10.1136/bmj.325.7370.961>.

154. Elizabeth A. McGlynn et al., *The Quality of Health Care Delivered to Adults in the United States*, 348 *NEJM* 2635, 2642–43 (2003), <https://doi.org/10.1056/NEJMsa022615>.

155. 2015 CDEHC Report, *supra* note 56, at 19.

156. Wennberg, *supra* note 153, at 962.

157. Michael J. Barry et al., *Patient Reactions to a Program Designed to Facilitate Patient Participation in Treatment Decisions for Benign Prostatic Hyperplasia*, 33 *Med. Care* 771, 778 (1995), <https://doi.org/10.1097/00005650-199508000-00003>.

supply-sensitive services refer to care that depends not on evidence of effectiveness or patient preferences but rather on the availability of services. Specifically, patients living in areas with more doctors or more hospitals experience more office visits, tests, and hospitalizations.¹⁵⁸

Evidence-Based Medicine

The variation in the delivery of medical care led to a careful reexamination of diagnostic strategies, therapeutic decision-making, and the use of medical evidence, but it was not the only factor. Other circumstances that set the stage for an intense focus on medical evidence included (1) the development of medical research, including randomized controlled trials and other observational study designs; (2) the growth of diagnostic and therapeutic interventions;¹⁵⁹ (3) interest in understanding medical decision-making and how physicians reason;¹⁶⁰ and (4) the acceptance of meta-analysis as a method to combine data from multiple randomized trials.¹⁶¹ In response to the above conditions, evidence-based medicine gained prominence in 1992,¹⁶² defined by Sackett et al. as “the conscientious, explicit and judicious use of current best evidence in making decisions about the care of the individual patient. It means integrating individual clinical expertise with the best available external clinical evidence from systematic research.”¹⁶³

Evidence-based medicine contrasts with the traditional informal method of practicing based on anecdotes, applying the most recently read articles, doing what a group of eminent experts recommend, or minimizing costs.¹⁶⁴ Rather,

158. Wennberg, *supra* note 153, at 962–63.

159. Cynthia D. Mulrow & K.N. Lohr, *Proof and Policy from Medical Research Evidence*, 26 J. Health Pol., Pol’y & L. 249, 250–56 (2001), <https://doi.org/10.1215/03616878-26-2-249>.

160. Robert S. Ledley & Lee B. Lusted, *Reasoning Foundations of Medical Diagnosis: Symbolic Logic, Probability, and Value Theory Aid Our Understanding of How Physicians Reason*, 130 Science 9, 9–10 (1959), <https://doi.org/10.1126/science.130.3366.9>.

161. See Steve C. Gold et al., *Reference Guide on Epidemiology*, “Methods for Synthesizing or Combining the Results of Multiple Studies,” in this manual; Video Software Dealers Ass’n v. Schwarzenegger, 556 F.3d 950, 963 (9th Cir. 2009) (analyzing a meta-analysis of studies on video games and adolescent behavior); Kennecott Greens Creek Min. Co. v. Mine Safety & Health Admin., 476 F.3d 946, 953 (D.C. Cir. 2007) (reviewing the Mine Safety and Health Administration’s reliance on epidemiological studies and two meta-analyses).

162. Evidence-Based Medicine Working Group, *Evidence-Based Medicine: A New Approach to Teaching the Practice of Medicine*, 268 JAMA 2420, 2420–21 (1992), <https://doi.org/10.1001/jama.1992.03490170092032>.

163. David L. Sackett et al., *Evidence Based Medicine: What It Is and What It Isn’t*, 312 B.M.J. 71, 71 (1996), <https://doi.org/10.1136/bmj.312.7023.71>.

164. Trisha Greenhalgh, *How to Read a Paper: The Basics of Evidence-Based Medicine* 5–9 (3d ed. 2006).

per Greenhalgh, it is “the use of mathematical estimates of the risks of benefit and harm, derived from high-quality research on population samples, to inform clinical decision making in the diagnosis, investigation or management of individual patients.”¹⁶⁵ In a paper from a joint workshop held by the IOM and the Agency for Healthcare Research and Quality that addressed what clinicians consider to be sufficient evidence to justify their clinical practice and treatment decisions, Mulrow and Lohr wrote, “evidence-based medicine stresses a structured critical examination of medical research literature; relatively speaking, it deemphasizes average practice as an adequate standard and personal heuristics.”¹⁶⁶

Hierarchy of Medical Evidence

With the explosion of available medical evidence, increased emphasis has been placed on assembling, evaluating, and interpreting medical research evidence. A fundamental principle of evidence-based medicine is that the strength of medical evidence supporting a therapy or strategy is hierarchical (see also section titled “Uncertainty and Tradeoffs in Therapeutic Decision-Making Evidence” below). When ordered from strongest to weakest, a systematic review of randomized trials (meta-analysis) is at the top, followed by single randomized trials, systematic reviews of observational studies, single observational studies, physiological studies, and unsystematic clinical observations.¹⁶⁷ An analysis of the frequency with which various study designs are cited by others provides empirical evidence supporting the influence of meta-analysis followed by randomized controlled trials in the medical evidence hierarchy.¹⁶⁸ Although they are at the bottom of the evidence hierarchy, unsystematic clinical observations or case reports may be the first signals of adverse events or associations that are later confirmed with larger or controlled epidemiological studies (e.g., aplastic anemia caused by chloramphenicol¹⁶⁹ or lung cancer caused by asbestos¹⁷⁰). Nonetheless, subsequent studies may not confirm initial reports (e.g., the putative

165. *Id.* at 1.

166. Mulrow & Lohr, *supra* note 159, at 253.

167. Gordon H. Guyatt et al., *Users’ Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice* 11 (2d ed. 2008); see also Steve C. Gold et al., *Reference Guide on Epidemiology*, “Methods for Synthesizing or Combining the Results of Multiple Studies,” in this manual.

168. Nikolaos A. Patsopoulos et al., *Relative Citation Impact of Various Study Designs in the Health Sciences*, 293 JAMA 2362, 2364–65 (2005), <https://doi.org/10.1001/jama.293.19.2362>.

169. W.T. Clarke, *Fatal Aplastic Anemia and Chloramphenicol*, 97 Can. Med. Ass’n J. 815 (1967), <https://perma.cc/9BR4-34H5>.

170. Michael Gochfeld, *Asbestos Exposure in Buildings*, in *Environmental Medicine* 438, 440 (Stuart M. Brooks et al. eds., 1995).

novel association between coffee consumption and pancreatic cancer published in a prominent medical journal).¹⁷¹

Just as basic science findings require repeated confirming experiments, evidence about the benefits and harms of medical interventions arise through repetitive observations. A single randomized controlled trial relies on hypothesis testing, specifically assuming the null hypothesis that a new drug is equivalent to the comparator (e.g., placebo). As conceived about 100 years ago, interpreting a trial involves calculating the likelihood of the alpha error (p -value) wherein the study suggests that the drug or device is beneficial, but the “truth” is that it is not, that is, a false-positive study result. Similarly, a beta error is the likelihood of a study finding that the drug or device is not beneficial when the “truth” is that it is, that is, a false-negative study result (1 minus power, where power is the likelihood of a study finding that the drug or device is not beneficial and the truth aligns with that finding).

The choice of which specific error rates to use (e.g., false-positive or p -value or alpha of 0.05) was supposed to depend on a judgment of the relative consequences of the two errors, missing an effective drug (Type II beta error) or considering an ineffective drug to be effective (Type I alpha error) with a stricter (lower) error rate for the latter.¹⁷² The null hypothesis, however, assumes equivalence, and so it does not provide any measure of evidence outside of the particular study (e.g., prior studies or biological mechanism or plausibility). Thus, the null hypothesis assumption necessitates abandoning the ability to measure evidence or determine “truth” from a single experiment so that hypothesis testing is thereby “equivalent to a system of justice that is not concerned with which individual defendant is found guilty or innocent (that is, ‘whether each separate hypothesis is true or false’) but tries instead to control the overall number of incorrect verdicts” over time in the long run.¹⁷³

Cumulative meta-analysis of treatments enables the accumulation of randomized trial evidence to examine trends in efficacy or risks, overcoming issues of underpowered trials with insufficient numbers of patients enrolled to reliably detect a benefit.¹⁷⁴ For example, between 1959 and 1988, 33 randomized trials with streptokinase for acute heart attack (myocardial infarction) involving over 35,000 patients were published. As each trial is published, it is combined with all previously published trials. A cumulative meta-analysis found “a consistent, statistically significant reduction in total mortality” $p < 0.001$ with streptokinase use by 1977, yet an additional 32,600 individuals

171. Brian MacMahon et al., *Coffee and Cancer of the Pancreas*, 304 NEJM 630, 631–32 (1981), <https://doi.org/10.1056/NEJM198103123041102>; Dominique S. Michaud et al., *Coffee and Alcohol Consumption and the Risk of Pancreatic Cancer in Two Prospective United States Cohorts*, 10 Cancer Epidemiology, Biomarkers & Prevention 429, 429 (2001).

172. Goodman, *supra* note 71, at 998.

173. *Id.*

174. Joseph Lau et al., *Cumulative Meta-Analysis of Therapeutic Trials for Myocardial Infarction*, 327 NEJM 248, 248 (1992), <https://doi.org/10.1056/NEJM199207233270406>.

underwent randomization through 1988, with half receiving the inferior placebo. In contrast, for many years, physicians used a prophylactic drug called lidocaine to prevent life-threatening heart-rhythm disturbances from acute heart attacks, yet none of the randomized trials of lidocaine demonstrated any benefit, and finally, the cumulative meta-analysis found a trend toward harm. When the results of the meta-analysis were compared with comments in textbooks and review articles,

discrepancies were detected between the meta-analytic patterns of effectiveness in the randomized trials and the recommendations of reviewers [the review article authors]. Review articles often failed to mention important advances or exhibited delays in recommending effective preventive measures. In some cases, treatments that have no effect on mortality or are potentially harmful continued to be recommended by several clinical experts.¹⁷⁵

Guidelines

Clinical-practice guidelines are “systematically developed statements to assist practitioner and patient decisions about appropriate healthcare for specific clinical circumstances.”¹⁷⁶ Such guidelines have been widely developed and issued by medical-specialty associations, professional societies, government agencies, and healthcare organizations.¹⁷⁷ To avoid biases inherent in review articles (particularly single-authored ones) and to encourage transparency and acceptance, the IOM recommended best practices for the development of clinical-practice guidelines, including process transparency, management of conflict of interest, guideline group composition (multidisciplinary and balanced), systematic reviews that meet IOM standards, evidence foundation for rating the strength of the recommendations, recommendation articulation, external review, and updating.¹⁷⁸

175. Elliott M. Antman et al., *A Comparison of Results of Meta-Analyses of Randomized Control Trials and Recommendations of Clinical Experts: Treatments for Myocardial Infarction*, 268 JAMA 240, 240 (1992).

176. 1990 CAPHSCPG Report, *supra* note 19, at 8.

177. See generally *Sofamor Danek Grp., Inc. v. Gaus*, 61 F.3d 929 (D.C. Cir. 1995) (reviewing guidelines issued by the Agency for Health Care Policy and Research in light of the Federal Advisory Committee Act); *Levine v. Rosen*, 616 A.2d 623 (Pa. 1992) (finding that differing guidance from two groups was evidence that reasonable physicians could follow either school of thought); Michelle M. Mello, *Of Swords and Shields: The Role of Clinical Practice Guidelines in Medical Malpractice Litigation*, 149 U. Pa. L. Rev. 645 (2001).

178. 2011 CSDTCPPG Report, *supra* note 20, at 5–9; Committee on Standards for Systematic Reviews of Comparative Effectiveness Research, Inst. Med., *Finding What Works in Health Care: Standards for Systematic Reviews* 6–11, 14–15 (Jill Eden et al. eds., 2011), <https://doi.org/10.17226/13059>.

The number, length, and diversity of guidelines developed by various professional organizations challenge practicing physicians.¹⁷⁹ Different professional organizations may issue guidelines on the same topic but with competing or similar recommendations. The composition of the panel and the processes for developing guideline recommendations may differ. For example, the United States Preventive Services Task Force (USPSTF) is “an independent panel of non-Federal experts in prevention and evidence-based medicine and is composed of primary care providers (such as internists, pediatricians, family physicians, gynecologists/obstetricians, nurses, and health behavior specialists).”¹⁸⁰ In their 2016 evaluation of mammography, the USPSTF “recommends biennial screening mammography for women aged 50 to 74 years” but “[t]he decision to start screening mammography in women prior to age 50 years should be an individual one.”¹⁸¹ In contrast, based on a writing group composed of radiologists, the American College of Radiology and Society of Breast Imaging recommend “annual mammography screening beginning at age 40” for all women at average risk.¹⁸²

For prostate cancer screening, however, the USPSTF’s 2018 update stated, “For men aged 55 to 69 years, the decision to undergo periodic PSA-based screening for prostate cancer should be an individual one and should include discussion of the potential benefits and harms of screening with their clinician.”¹⁸³ In the 2013 American Urological Association update, a statement panel composed of urologists, primary care physicians, radiation and medical oncologists, and epidemiologists “recommended shared decision-making for men age 55 to 69 years considering PSA-based screening, a target age group for whom benefits may outweigh harms”—essentially identical recommendations.

Practice guidelines provide recommendations on how to evaluate and treat patients, but because they apply to the general case, their recommendations may not apply to a particular individual patient—or some extrapolation may be required, particularly when multiple diseases exist, as in the elderly,¹⁸⁴ or when

179. Arthur Hibble et al., *Guidelines in General Practice: The New Tower of Babel?*, 317 B.M.J. 862, 862 (1998), <https://doi.org/10.1136/bmj.317.7162.862>.

180. U.S. Preventive Servs. Task Force (USPSTF), Agency for Healthcare Rsch. and Quality, <https://perma.cc/ZDT5-X752>.

181. Albert L. Siu et al., *Screening for Breast Cancer: U.S. Preventive Services Task Force Recommendation Statement*, 164 *Annals Internal Med.* 279, 279, 282–83 (2016), <https://doi.org/10.7326/M15-2886>; see also section titled “When to Start Screening” in the report.

182. Debra L. Monticciolo et al., *Breast Cancer Screening Recommendations Inclusive of All Women at Average Risk: Update from the ACR and Society of Breast Imaging*, 18 *J. Am. Coll. Radiology* 1280, 1280 (2021), <https://doi.org/10.1016/j.jacr.2021.04.021>.

183. U.S. Preventive Servs. Task Force et al., *Screening for Prostate Cancer: U.S. Preventive Services Task Force Recommendation Statement*, 319 *JAMA* 1901, 1901 (2018), <https://doi.org/10.1001/jama.2018.3710>.

184. Cynthia M. Boyd et al., *Clinical Practice Guidelines and Quality of Care for Older Patients with Multiple Comorbid Diseases: Implications for Pay for Performance*, 294 *JAMA* 716, 718–22 (2005),

treatment entails competing risks of harm. For example, anticoagulation is generally recommended for patients with atrial fibrillation (an abnormal heart rhythm disturbance) to prevent blood clots in the heart that could flow to the brain and cause a stroke, yet anticoagulation can also lead to life-threatening bleeding; therefore, for individual patients, physicians must weigh the risk of developing clots versus the risk of bleeding. Consequently, guidelines typically include statements such as “clinical or policy decisions involve more considerations than this body of evidence alone. Clinicians and policymakers should understand the evidence but individualize decision-making to the specific patient or situation.”¹⁸⁵

Some physicians who rely on personal style, review articles, and colleagues to influence their clinical practice have been concerned with how guidelines affect clinical autonomy and healthcare costs.¹⁸⁶ But just as clinicians have been reluctant to apply guidelines in practice, courts have generally been slow to apply them in deciding cases.¹⁸⁷ Political and legal issues have arisen with the development of guidelines.¹⁸⁸ In 2006, the Connecticut attorney general launched an antitrust suit against the Infectious Disease Society of America (IDSA) after IDSA promulgated a guideline recommending against the use of long-term antibiotics for the treatment of chronic Lyme disease.¹⁸⁹ Although the findings of the Centers for Disease Control and Prevention and the Food and Drug Administration (FDA) seemed to concur with IDSA’s guidelines, an advocacy group representing patients afflicted with chronic Lyme disease and the physicians who treated them protested; the attorney general’s decision to investigate followed shortly afterwards.¹⁹⁰ Organizations can violate antitrust laws if their guideline-setting process is an unreasonable attempt to advance their members’ economic interests by suppressing competition. After legal expenses that exceeded \$250,000, IDSA settled without admitting any fault, providing an example of

<https://doi.org/10.1001/jama.294.6.716>.

185. U.S. Preventive Services Task Force, *Screening for Carotid Artery Stenosis: U.S. Preventive Services Task Force Recommendation Statement*, 147 *Annals Internal Med.* 854, 854 (2007), <https://doi.org/10.7326/0003-4819-147-12-200712180-00005>.

186. Sean R. Tunis et al., *Internists’ Attitudes About Clinical Practice Guidelines*, 120 *Annals Internal Med.* 956, 956 (1994), <https://doi.org/10.7326/0003-4819-120-11-199406010-00008>.

187. Arnold J. Rosoff, *Evidence-Based Medicine and the Law: The Courts Confront Clinical Practice Guidelines*, 26 *J. Health Pol., Pol’y & L.* 327, 330–53 (2001), <https://doi.org/10.1215/03616878-26-2-327>.

188. One element in the near demise of the Agency for Health Care Policy and Research (AHCPR) was a political audience receptive to complaints from an association of back surgeons who disagreed with the AHCPR practice guideline conclusions regarding low back pain. B.H. Gray et al., *AHCPR and the Changing Politics of Health Services Research*, 22 *Health Affs. (Supp. 1)* W3-283–307 (2003), <https://doi.org/10.1377/hlthaff.w3.283>.

189. John D. Kraemer & Lawrence O. Gostin, *Science, Politics, and Values: The Politicization of Professional Practice Guidelines*, 301 *JAMA* 665, 665 (2009), <https://doi.org/10.1001/jama.201.6.665>.

190. *Id.* at 666.

“the politicization of health policy, with elected officials advocating for health policies against the weight of scientific evidence.”¹⁹¹

Besides clinical practice guidelines, the IOM defined other types of statements: (1) medical review criteria are systematically developed statements that can be used to assess the appropriateness of specific healthcare decisions, services, and outcomes; (2) standards of quality are authoritative statements of minimum levels of acceptable performance or results, excellent levels of performance or results, or the range of acceptable performance or results; and (3) performance measures are methods or instruments to estimate or monitor the extent to which the actions of a healthcare practitioner or provider conform to practice guidelines, medical review criteria, or standards of quality.¹⁹²

Uncertainty and Tradeoffs in Therapeutic Decision-Making Evidence

Medical decision-making often involves complexity, uncertainty, and tradeoffs¹⁹³ because of unique genetic factors, lifestyle habits, known conditions, medication histories, and ambiguity about possible diagnoses, test results, treatment benefits, and therapeutic harms. Given inherent diagnostic and therapeutic uncertainty, clinicians often make treatment decisions in the face of uncertainty.

Well-performed randomized trials provide the least biased estimates of treatment benefit and harm by comparing groups with equivalent prognoses. Sticking strictly to the scientific evidence, some physicians may limit their use of medications to the specific drug at the specific doses found to be beneficial in such trials. Others may assume class effects (drugs with similar structure, mechanism of action, or pharmacologic effects, e.g., lowering blood pressure) until proven otherwise. Others may consider additional factors such as out-of-pocket costs for patients or patient preferences. When physicians evaluate patients who might benefit from a treatment but who would have been excluded from the study in which the benefit was demonstrated, they must weigh the risks of harm and benefits through extrapolation in the absence of definitive evidence of benefit or of harm. Indeed, because few medical recommendations are based on randomized trials (the least biased level of evidence), physicians frequently and necessarily face uncertainty in making testing and treatment decisions. For example, in an academic pediatric cardiology hospital, only 3% of clinically

191. *Id.* at 665.

192. 1990 CAPHSCPG Report, *supra* note 19, at 8.

193. John P. A. Ioannidis & Joseph Lau, *Systematic Review of Medical Evidence*, 12 J. L. & Pol’y 509, 511, 524 (2004).

significant decisions are related to a specific study.¹⁹⁴ In most disciplines, clear evidence of efficacy and risks of treatment are lacking. In cardiology (one of the best-studied areas of medical care), nearly one-half of guideline recommendations are based on expert opinion, case studies, or standards of care.¹⁹⁵

Applying well-designed studies to populations of patients represents another problem. The Randomized Aldactone Evaluation Study demonstrated that spironolactone reduced mortality and hospitalizations for heart failure and improved quality of life with minimal risk of seriously high levels of potassium (hyperkalemia).¹⁹⁶ Published in a prominent medical journal, prescriptions for spironolactone rose quickly because of familiarity with the medication and the poor prognosis of patients with heart failure. In Ontario, hospitalizations per 1,000 patients for high potassium, however, rose from 2.4 in 1994 to 11.0 in 2001, resulting in an estimated 560 additional hospitalizations for high potassium and 73 additional hospital deaths in older patients with heart failure.¹⁹⁷ As opposed to the study population, community individuals were older, and more frequently women, and more often had absolute or relative contraindications to treatment with only 25% characterized as ideal candidates for spironolactone.¹⁹⁸ These factors increased the risk that spironolactone therapy would lead to high potassium levels that could be life-threatening. Criteria for entry into randomized trials of drugs typically exclude individuals with concomitant medication use, with medical comorbidities, who are female, and who have disadvantaged socioeconomic status, thereby limiting the ability to generalize the results of a trial to the clinical population being treated.¹⁹⁹ Clinicians refer to randomized controlled studies as assessments of drug efficacy in restricted patient populations, whereas studies of treatment in general clinical populations are often referred to as effectiveness studies.

To be sufficiently powered (i.e., to have enough patients enrolled to avoid Type II error) to demonstrate statistical significance,²⁰⁰ randomized controlled trials

194. Jeffrey R. Darst et al., *Deciding Without Data*, 5 *Congenital Heart Disease* 339, 340 (2010), <https://doi.org/10.1111/j.1747-0803-2010.00433.x>.

195. Pierluigi Tricoci et al., *Scientific Evidence Underlying the ACC/AHA Clinical Practice Guidelines*, 301 *JAMA* 831, 835 (2009), <https://doi.org/10.1001/jama.2009.205>.

196. Bertram Pitt et al., *The Effect of Spironolactone on Morbidity and Mortality in Patients with Severe Heart Failure: Randomized Aldactone Evaluation Study Investigators*, 341 *NEJM* 709, 709 (1999), <https://doi.org/10.1056/NEJM199909023411001>.

197. David N. Juurlink et al., *Rates of Hyperkalemia After Publication of the Randomized Aldactone Evaluation Study*, 351 *NEJM* 543, 543 (2004), <https://doi.org/10.1056/NEJMoa040135>.

198. Dennis T. Ko et al., *Appropriateness of Spironolactone Prescribing in Heart Failure Patients: A Population-Based Study*, 12 *J. Cardiac Failure* 205, 206 (2006), <https://doi.org/10.1016/j.cardfail.2006.01.003>.

199. Harriette G. C. Van Spall et al., *Eligibility Criteria of Randomized Controlled Trials Published in High-Impact General Medical Journals: A Systematic Sampling Review*, 297 *JAMA* 1233, 1237 (2007), <https://doi.org/10.1001/jama.297.11.1233>.

200. See Steve C. Gold et al., *Reference Guide on Epidemiology*, “Methods for Synthesizing or Combining the Results of Multiple Studies,” in this manual.

usually require high event rates, prolonged follow-up, or large numbers of patients. Because of impracticality, expense, and the time period needed to obtain long-term outcomes, these trials may often choose a surrogate marker that is associated with a clinically important event or with survival. For example, statins were approved on the basis of their safety and efficacy in lowering cholesterol but were only demonstrated to improve survival in patients with known coronary heart disease years later.²⁰¹ Fast-track approval of new drugs for HIV infection was based on safety and efficacy in reducing viral levels (as a surrogate or substitute outcome measure felt to be related to survival) as opposed to a demonstration of improved survival.

On the other hand, in the late 1970s, patients with frequent extra heartbeats (ventricular premature contractions) following a heart attack had an increased risk of sudden death. On that basis, those in the then-emerging field of cardiac electrophysiology believed that reducing ventricular premature beats (as a surrogate outcome measure) would decrease subsequent sudden cardiac death. In early randomized controlled trials, oral antiarrhythmic drugs such as encainide and flecainide were approved by the FDA on the basis of their ability to suppress these extra heartbeats in patients who had had a myocardial infarction. Years after approval of these drugs, however, a randomized controlled trial designed to demonstrate the survival benefit of these drugs was discontinued after only ten months because of a statistically significant higher rate of mortality in patients receiving the drugs. Although these drugs effectively suppressed the extra heartbeats, the study found that they actually increased the likelihood of fatal heart rhythm disturbances.²⁰²

Prior to approval by the FDA, drugs and devices must undergo Phase 1, 2, and 3 clinical trials to demonstrate safety and efficacy. Following preliminary chemical discovery, toxicology, and animal studies, Phase 1 studies examine the safety of new drugs in healthy individuals. Phase 2 studies involve varying drug doses in individuals with the disease to explore efficacy and responses and adverse effects. Based on the dose or doses identified in Phase 2, a Phase 3 study examines drug response in a larger number of patients to again determine safety and efficacy in the hope of getting a new drug approved for sale by regulatory authorities. However, because fewer than 10,000 individuals have usually received the drug during most trials, uncommon adverse outcomes may not become apparent until usage is broadened and extended. For example, about 1 in 24,200, or

201. *Randomised Trial of Cholesterol Lowering in 4444 Patients with Coronary Heart Disease: The Scandinavian Simvastatin Survival Study (4S)*, 344 *Lancet* 1383, 1385 (1994), [https://doi.org/10.1016/S0140-6736\(94\)90566-5](https://doi.org/10.1016/S0140-6736(94)90566-5).

202. *Preliminary Report: Effect of Encainide and Flecainide on Mortality in a Randomized Trial of Arrhythmia Suppression After Myocardial Infarction: The Cardiac Arrhythmia Suppression Trial (CAST) Investigators*, 321 *NEJM* 406, 409 (1989), <https://doi.org/10.1056/NEJM198908103210629>.

40,500, patients who received the antibiotic chloramphenicol²⁰³ developed fatal aplastic anemia (in which the bone marrow no longer produces any blood cells). This adverse effect was discovered only in the 1960s after chloramphenicol was initially considered safe and had been widely used during the 1950s.²⁰⁴

For all approved drug and therapeutic biological products, the FDA managed postmarketing safety surveillance beginning in 1969 through the Adverse Event Reporting System, a voluntary reporting system with many limitations. For example, in 1999, rofecoxib (Vioxx), a Cox-2 selective nonsteroidal anti-inflammatory drug (NSAID), was approved for pain relief in part on the basis of studies that suggested that it induced less gastrointestinal bleeding than other NSAIDs. In 2004, the manufacturer announced a voluntary worldwide withdrawal of rofecoxib when a prospective study confirmed that the drug increased the risk of myocardial infarctions (heart attacks) and stroke with chronic use.²⁰⁵ Beginning in 2014 and officially in 2016, the FDA began Active Postmarket Risk Identification and Analysis (ARIA) with the development of the Sentinel Initiative to analyze real-world safety data through “the largest multisite distributed database in the world dedicated to medical product safety” that involves big data from twenty-five academic, health-system, and health-insurer collaborators with experience and expertise in epidemiology, clinical medicine, pharmacy, statistics, health informatics, and data science (natural language processing and machine learning).²⁰⁶

Even in a randomized trial in which a drug is found to be beneficial, some patients who received the drug may have been harmed, emphasizing the need to individualize the balancing of risks of harm and benefits and explaining in part why some clinicians may not adhere to guideline recommendations. The fundamental dilemma was articulated by Bernard in 1865: The response of the “average” patient to therapy is not necessarily the response of the patient being treated.²⁰⁷ Indeed, the average results of clinical trials do not apply to all patients in the trial; this is called heterogeneity of treatment effects²⁰⁸ wherein even with well-defined inclusion and exclusion criteria, variation in outcome risk, and therefore treatment benefit, exists so that even “typical” patients included in the trial may not get the average observed trial benefits.

The Global Utilization of Streptokinase and tPA for Occluded Coronary Arteries Trial is a case in point. The trial suggested that accelerated tissue

203. Clarke, *supra* note 169, at 815.

204. *Id.*

205. *See generally* *In re Vioxx Prods. Liab. Litig.*, 360 F. Supp. 2d 1352 (J.P.M.L. 2005).

206. U.S. Food & Drug Administration, FDA’s Sentinel Initiative—Background, <https://www.fda.gov/safety/fdas-sentinel-initiative>.

207. Salim Yusuf et al., *Analysis and Interpretation of Treatment Effects in Subgroups of Patients in Randomized Clinical Trials*, 266 JAMA 93, 93 (1991), <https://doi.org/10.1001/jama.1991.03470010097038>.

208. David M. Kent et al., *The Predictive Approaches to Treatment Effect Heterogeneity (PATH) Statement*, 172 Annals Internal Med. 35, 41 (2020), <https://doi.org/10.7326/M18-3667>.

plasminogen accelerator (tPA), a potent anticoagulant, reduced mortality from acute heart attacks (myocardial infarction) by dissolving blood clots, with the tradeoff being an increased risk of bleeding in the brain from tPA.²⁰⁹ In a reanalysis of this study, most (85%) of the survival benefit of tPA accrued to half of the patients (those at highest risk of dying from their heart attack). Some patients with very low risk of dying from their heart attack who received tPA were likely harmed because their risk of bleeding in their brain exceeded the benefit.²¹⁰ Therefore, to optimize treatment decisions in practice, clinicians may attempt to individualize treatment decisions based on their assessment of the patient's risk of harm versus benefit. Even then, clinicians may be reluctant to administer a medication such as tPA that can cause severe harm, such as bleeding into the brain (intracranial hemorrhage). A single clinical experience with a patient who bled when given tPA might well color a clinician's judgment about the benefits of the treatment going forward (the availability heuristic).

A fundamental principle of evidence-based medicine is that “[e]vidence alone is never sufficient to make a clinical decision.”²¹¹ Nearly all medical decisions involve some tradeoff between a benefit and a harm. Besides the options and the likelihood of the outcomes, patient preferences about the resulting outcomes should affect care choices, especially when there are tradeoffs, such as a risk of complications or dying from a procedure or treatment versus some benefit such as living longer (provided the patient survives the short-term risk of the procedure) or improving their quality of life (relieving symptoms). Besides individualizing risk of harm and benefit assessments, clinicians may also deviate from guideline recommendations (warranted variation) because of a particular patient's higher risk of adverse events or lower likelihood of benefit or because of patient preferences for the alternative outcomes, such as when risks occur at different times. For example, given a hypothetical choice between living 12.5 years for certain or a 50:50 chance of living 25 years or dying immediately, most individuals would choose 12.5 years. Although both options yield, on average, 12.5 years, individuals tend to be risk averse and prefer to avoid the near-term risk of dying. When interviewed, some patients with operable lung cancer were quite averse to possible immediate death from surgery, and so, based on their

209. The GUSTO Investigators, *An International Randomized Trial Comparing Four Thrombolytic Strategies for Acute Myocardial Infarction*, 329 *NEJM* 673, 673 (1993), <https://doi.org/10.1056/NEJM199309023291001>.

210. David M. Kent et al., *An Independently Derived and Validated Predictive Model for Selecting Patients with Myocardial Infarction Who Are Likely to Benefit from Tissue Plasminogen Activator Compared with Streptokinase*, 113 *Am. J. Med.* 104, 108 (2002), [https://doi.org/10.1016/s0002-9343\(02\)01160-9](https://doi.org/10.1016/s0002-9343(02)01160-9).

211. Guyatt et al., *supra* note 167, at 8; see also section titled “Hierarchy of Medical Evidence” above.

preferences, these patients probably should opt for radiation therapy despite its poorer long-term survival.²¹²

Besides risk aversion, some treatments may improve quality of life but place patients at risk for shortened life expectancy, and some patients may be willing to trade off quality of life for length of life. When presented with laryngeal cancer scenarios, some volunteer research subjects chose radiation therapy over surgery to preserve their voices despite a reduced likelihood of future survival. “These results suggest that treatment choices should be made on the basis of patients’ attitudes toward the quality as well as the quantity of survival.”²¹³

To illustrate this principle, a National Institutes of Health Consensus Conference recommended breast-conserving surgery when possible for women with Stage I and II breast cancer, because well-designed studies with long-term follow-up on thousands of women demonstrated equivalence of lumpectomy and radiation therapy or mastectomy for survival and disease-free survival (being alive without breast cancer recurrence).²¹⁴ In one study, lumpectomy and radiation appeared to have a lower risk of breast cancer recurrence, with five women reported to have had breast cancer recurrences following lumpectomy and radiation versus ten women after mastectomy.²¹⁵ However, breast cancer that recurred in the breast that had been operated on was censored (i.e., deliberately not considered in the statistical analysis, likely because the breast would have been removed with the alternative intervention mastectomy).²¹⁶ When including these censored cancer recurrences, 20 breast cancer recurrences occurred after lumpectomy versus ten after mastectomy, and so lumpectomy actually had a higher overall risk of recurrence.²¹⁷ As expressed by one woman, “The decision about treatment for breast cancer remains an intensely personal one. The mastectomy I chose . . . felt a lot less invasive than the prospect of six weeks of daily radiation, not to mention the 14% risk of local recurrence.”²¹⁸ In such a case,

212. Barbara J. McNeil et al., *Speech and Survival: Tradeoffs Between Quality and Quantity of Life in Laryngeal Cancer*, 305 NEJM 982, 986 (1981), <https://doi.org/10.1056/NEJM198110223051704>.

213. *Id.* at 982.

214. *NIH Consensus Conference: Treatment of Early-Stage Breast Cancer*, 265 JAMA 391, 392 (1991), <https://doi.org/10.1001/jama.1991.03460030097037>.

215. Joan A. Jacobson et al., *Ten-Year Results of a Comparison of Conservation with Mastectomy in the Treatment of Stage I and II Breast Cancer*, 332 NEJM 907, 909 (1995), <https://doi.org/10.1056/NEJM199504063321402>.

216. Bernard Fisher et al., *Eight-Year Results of a Randomized Clinical Trial Comparing Total Mastectomy and Lumpectomy With or Without Irradiation in the Treatment of Breast Cancer*, 320 NEJM 822, 824 (1989), <https://doi.org/10.1056/NEJM198903003201302>; Jacobson et al., *supra* note 215, at 909.

217. Jacobson et al., *supra* note 215, at 909.

218. Karen Sepucha et al., *Policy Support for Patient-Centered Care: The Need for Measurable Improvements in Decision Quality*, 23 Health Affs. (Supp. 2) VAR54, VAR62 (2004), <https://doi.org/10.1377/hlthaff.var.54>.

patient preferences²¹⁹ about tradeoffs involving breast preservation and increased risk of breast cancer recurrence or the need for radiation therapy associated with lumpectomy may play an important role in determining the optimal decision for any particular patient.²²⁰

Biases

Cognitive psychology

In *Predictably Irrational: The Hidden Forces That Shape Our Decisions*, Dan Ariely refers to

the simple and compelling idea that we are capable of making the right decision for ourselves. . . . In fact, this book is about human *irrationality*—about our distance from perfection. . . . Understanding irrationality is important for our everyday actions and decisions, and for understanding how we design our environment and the choices it presents to us.²²¹

Beyond the cognitive biases associated with heuristics (representativeness, availability, and anchoring, mentioned earlier) and patient- and hospital-related factors leading to medical errors, medical decision-making is susceptible to pitfalls related to other cognitive biases, including framing effects (being influenced by wording and context, e.g., choosing riskier options when described in negative terms such as mortality), blind obedience (deference to authority or technology), number of alternatives (choosing one treatment more often when presented with additional options), premature closure (focusing narrowly on a single opinion), and personality traits such as tolerance of ambiguity or overconfidence.²²² A systematic review of cognitive biases in medical decisions made by physicians found that at least one cognitive factor or bias was present in every study, with five of seven studies showing an association with medical errors.²²³ Lastly, an

219. Proctor & Gamble Pharm., Inc. v. Hoffman-LaRoche, Inc., No. 06 CIV. 0034 (PAC), 2006 WL 2588002, at *10 (S.D.N.Y. Sept. 6, 2006) (detailing the testimony of a physician stating that, in addition to efficacy, he considers patient preferences when determining treatment for osteoporosis).

220. Jerome P. Kassirer, *Adding Insult to Injury: Usurping Patients' Prerogatives*, 308 NEJM 898, 900–01 (1983), <https://doi.org/10.1056/NEJM198304143081511>.

221. Dan Ariely, *Predictably Irrational: The Hidden Forces That Shape Our Decisions* xix (2010).

222. Gustavo Saposnik et al., *Cognitive Biases Associated with Medical Decisions: A Systematic Review*, 16 BMC Med. Informatics & Decision Making 138 (2016), <https://doi.org/10.1186/s12911-016-0377-1>; Donald A. Redelmeier, *Improving Patient Care: The Cognitive Psychology of Missed Diagnoses*, 142 Annals Internal Med. 115, 119 (2005), <https://doi.org/10.7326/0003-4819-142-2-200501180-00010>.

223. Saposnik et al., *supra* note 222, at 5.

emerging area concerns noise, that is, variability in judgments that should be identical. Specifically, different physicians do not make identical decisions for the same patient; this noise was especially high in psychiatry but found even in radiology (with, among other examples, forensic science and bail decisions): “Whenever you look at human judgments, you are likely to find noise [random scatter]. To improve the quality of our judgments, we need to overcome noise as well as bias [systematic deviation].”²²⁴

Race and ethnicity

Beyond the biases associated with the types and severity of the diseases that physicians may see, the 2003 IOM report *Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care* found that “[r]acial and ethnic minorities tend to receive a lower quality of healthcare than non-minorities, even when access-related factors, such as patients’ insurance status and income, are controlled. The sources of these disparities are complex, are rooted in historic and contemporary inequities, and involve many participants at several levels, including health systems, their administrative and bureaucratic processes, utilization managers, healthcare professionals, and patients.”²²⁵ Evidence of barriers leading to unequal treatment included access (even when equally insured), language, geography, cultural familiarity, health-system financial and institutional arrangements, and legal, regulatory, and policy environments.²²⁶ These health inequities arose from a “historic context in which healthcare has been differentially allocated on the basis of social class, race, and ethnicity.”²²⁷ Although this report is more than 20 years old, a 2021 study examining screening colonoscopy found consistent evidence of inequities including access to screening, the quality of screening, delay from diagnosis to initiation of treatment, and quality of treatment.²²⁸ Thus, criticism of the inclusion of race in prediction models (referred to as clinical algorithms) arose in major journals because “medicine is not a stand-alone institution immune to racial inequities, but rather is an institution of structural racism. A pervasive example of this participation is race-based medicine, the system by which research characterizing race as an

224. Daniel Kahneman et al., Noise: A Flaw in Human Judgment 4, 7, 273–86 (2021).

225. Comm. on Understanding and Eliminating Racial and Ethnic Disparities in Health Care, Inst. Med., *Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care* 1, 1 (2003), <https://doi.org/10.17226/12875>.

226. *Id.* at 140–59.

227. *Id.* at 123.

228. U.S. Preventive Services Task Force, *Screening for Colorectal Cancer: US Preventive Services Task Force Recommendation Statement*, 325 JAMA 1965, 1970 (2021), <https://doi.org/10.1001/jama.2021.6238>; Carolyn M. Rutter, *Black and White Differences in Colorectal Cancer Screening and Screening Outcomes: A Narrative Review*, 30 Cancer Epidemiology, Biomarkers & Prevention 3, 9–10 (2021), <https://doi.org/10.1158/1055-9965.EPI-19-1537>.

essential biological variable, translates into clinical practice, leading to inequitable care”;²²⁹ “[t]hese algorithms propagate race-based medicine”;²³⁰ and “[r]ace is a historically derived social construct that has no place as a biological proxy.”²³¹

Estimating kidney function with race and without

One area of focus became the estimation of kidney function and reconsidering whether the equation for estimating the glomerular filtration rate (GFR) should include race. The inclusion of race historically led to higher values of kidney function for people who are black. Those higher values could avoid overdiagnosis and overtreatment but could also delay referral to specialty care and transplantation and result in poorer health outcomes, particularly in a population at higher risk for end-stage kidney disease and related mortality than the general population.²³² Others have raised performance and data issues including validation, overall and within-group accuracy, availability of assays, and equation parameters in patients who are black.²³³

In response, the National Kidney Foundation and American Society of Nephrology created the NKF-ASN Task Force on Reassessing the Inclusion of Race in Diagnosing Kidney Disease. The task force, including ninety-seven experts with input from the community at large (students, trainees, clinicians, scientists, other health professionals, patients, family members and other public stakeholders), clarified the problem and evidence, narrowed the list of twenty-six approaches that did or did not consider race for estimating GFR to five, and evaluated those on “assay availability and standardization; implementation; population diversity in equation development; performance compared to measured GFR; consequences to clinical care, population tracking and research; and patient centeredness.”²³⁴ They ultimately recommended immediate implementation of a CKD-EPI (Chronic Kidney Disease Epidemiology Collaboration) creatinine equation that was revised by excluding the race variable in calculation and reporting based on “acceptable performance characteristics

229. Jessica P. Cerdeña et al., *From Race-Based to Race-Conscious Medicine: How Anti-Racist Uprisings Call Us to Act*, 396 *Lancet* 1125, 1125 (2020), [https://doi.org/10.1016/S0140-6736\(20\)32076-6](https://doi.org/10.1016/S0140-6736(20)32076-6).

230. Darshali A. Vyas et al., *Hidden in Plain Sight: Reconsidering the Use of Race Correction in Clinical Algorithms*, 383 *NEJM* 874, 874 (2020), <https://doi.org/10.1056/NEJMms2004740>.

231. Joseph L. Wright et al., *Eliminating Race-Based Medicine*, 150 *Pediatrics* e2022057998, 1 (2022), <https://doi.org/10.1542/peds.2022-057998>.

232. Vyas et al., *supra* note 230, at 875.

233. James A. Diao et al., *In Search of a Better Equation: Performance and Equity in Estimates of Kidney Function*, 384 *NEJM* 396, 398 (2021), <https://doi.org/10.1056/NEJMp2028243>.

234. Cynthia Delgado et al., *A Unifying Approach for GFR Estimation: Recommendations of the NKF-ASN Task Force on Reassessing the Inclusion of Race in Diagnosing Kidney Disease*, 32 *J. Am. Soc’y Nephrology* 2994, 2994 (2021), <https://doi.org/10.1681/ASN.2021070988>.

and potential consequences that do not disproportionately affect any one group of individuals.”²³⁵

Algorithms as standards

Beyond estimating kidney function, prediction algorithms and risk scores have become embedded in some guidelines to help clinicians to individualize treatment decisions—for example, for recommending a statin based on estimating a 10-year risk of having a stroke or heart attack with the American College of Cardiology/American Heart Association (ACC/AHA) Pooled Cohort Equations.²³⁶ However, studies suggest that these equations overpredict (higher estimates than those observed) in broad populations and underpredict (lower estimates than those observed) in disadvantaged communities. Moreover, the ACC/AHA, the U.S. Preventive Services Task Force, and the U.S. Department of Veterans Affairs each have different ten-year thresholds of heart attack or stroke for initiating a statin ranging from 7.5% to 12% or greater.²³⁷

A 2023 publication promoted using an algorithm to provide individualized risk for lung cancer. The editorial to that publication, however, points out that the use of that algorithm would have implications on safety (improved effectiveness of patient selection but worse safety profile, i.e., unnecessary invasive procedures and treatment of overdiagnosed cancers), patient-centeredness (the unmet need for clarity about patient values, preferences, and goals), timeliness (delays due to incorporating the algorithm into the electronic health record and an altered workflow), and equity (if minority-serving institutions have fewer resources to implement).²³⁸ On this topic the U.S. Preventive Services Task Force identifies the need for research “on the benefits and harms of using risk prediction models to select patients for lung cancer screening, including whether use of risk prediction models represents a barrier to wider implementation of lung cancer screening in primary care.” The task force recommended annual lung cancer screening for those aged 50 to 80 years old with a measure of smoking exposure (at least 20 pack-years, e.g., one pack of cigarettes smoked daily for 20 years), and with less than 15 years elapsed since stopping smoking.²³⁹

235. *Id.*

236. U.S. Preventive Servs. Task Force, *Statin Use for the Primary Prevention of Cardiovascular Disease in Adults: U.S. Preventive Services Task Force Recommendation Statement*, 328 JAMA 746, 747 (2022), <https://doi.org/10.1001/jama.2022.13044>.

237. *Id.* at 751–52.

238. Renda Soylemez Wiener & Michael K. Gould, *Selecting Candidates for Lung Cancer Screening: Implications for Effectiveness, Efficiency, Equity, and Implementation*, 176 Annals Internal Med. 413, 413–14 (2023), <https://doi.org/10.7326/M23-0230>.

239. U.S. Preventive Servs. Task Force, *Screening for Lung Cancer: U.S. Preventive Services Task Force Recommendation Statement*, 325 JAMA 962, 968 (2021), <https://doi.org/10.1001/jama.2021.1117>.

Thus, although a few clinical prediction or risk scores are recommended or mentioned in some guidelines and in use to assist with risk prediction and treatment decisions, implementation, external validity, and other issues as described above remain.

Algorithmic fairness and clinical prediction

In a perspective titled *Predictably Unequal: Understanding and Addressing Concerns That Algorithmic Clinical Prediction May Increase Health Disparities*, Paulus and Kent discuss concepts of algorithmic fairness versus algorithmic bias in healthcare in the context of increasing use of predictive algorithms (e.g., machine learning, artificial intelligence, or prediction models) to support decision-making.²⁴⁰ They acknowledge that fairness concerns apply when these predictive model algorithms lead to polar decisions, that is, where the predictions lead to decisions that benefit some groups but not others, such as in the allocation of healthcare resources, particularly scarce ones.²⁴¹ They examine different fairness criteria and demonstrate that “[e]ven when models are used to balance benefits–harms to make optimal decisions for individuals (i.e., for non-polar decisions)—and fairness concerns are not germane—model, data or sampling issues can lead to biased predictions that support decisions that are differentially harmful/beneficial across groups.”²⁴² Although they review potential sources of bias and methods for diagnosing and remedying bias, the core problematic issue may be that we lack agreed-upon definitions of fairness.

For example, they cite a commercial software used to predict reoffense after release based on data from 7,000 arrests.²⁴³ It found a higher risk of reoffense in defendants who were black, leading to potentially longer sentences, than those who were white, even though the algorithm was “race-unaware.” This was even the case among those who did not recidivate, that is, truly zero risk, resulting in unequal error rates consistent with “unfairness” and perhaps also “legal definitions of discrimination through ‘disparate impact.’” The software developers, however, argued that the model had good calibration—that is, good agreement between the observed and the predicted reoffense. However, when examining two fairness criteria that equalized either (1) type I/II error rates (sensitivity and specificity) or (2) test fairness with consistent calibration (strata-specific outcome

240. Jessica K. Paulus & David M. Kent, *Predictably Unequal: Understanding and Addressing Concerns That Algorithmic Clinical Prediction May Increase Health Disparities*, 3 N.P.J. Digit. Med. 99, at 1 (2020), <https://perma.cc/W2NM-KH3A>. See also James E. Baker & Laurie N. Hobart, *Reference Guide on Artificial Intelligence*, “Bias,” in this manual.

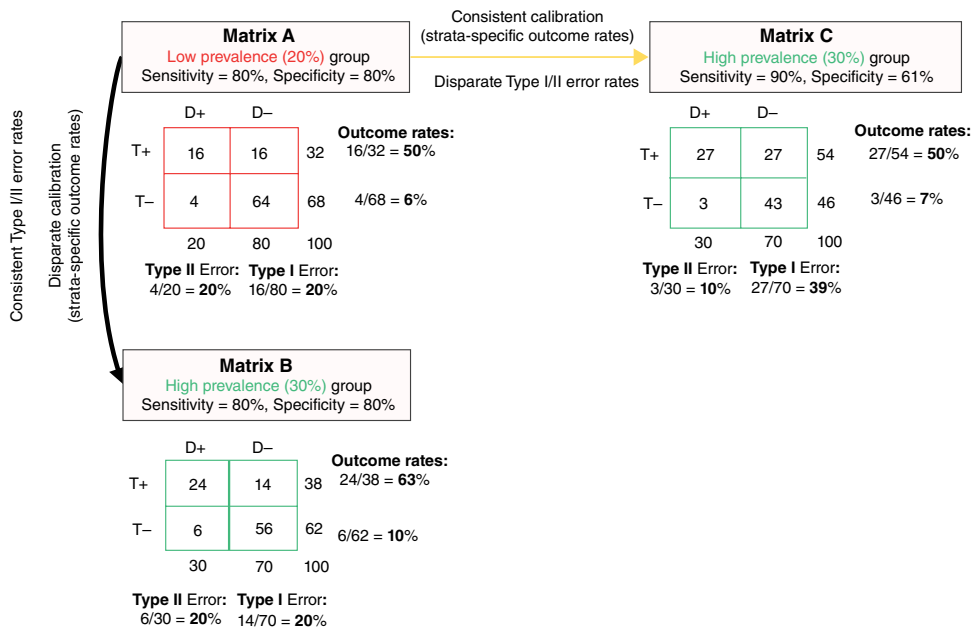
241. Paulus & Kent, *supra* note 240, at 1.

242. *Id.*

243. *Id.* at 2. See also Valena E. Beety et al., *Reference Guide on Forensic Feature Comparison Evidence*, “Special Issues with Machine-Generated Feature Comparison Evidence,” in this manual.

rates), Paulus and Kent found both criteria cannot simultaneously be satisfied when the outcome rates differ across two groups (except in the case of perfect prediction), “leading to the conclusion that unfairness is inevitable” (Figure 6).²⁴⁴

Figure 6. Mutual incompatibility of fairness criteria.



Source: From Jessica K. Paulus & David M. Kent, *Predictably Unequal: Understanding and Addressing Concerns That Algorithmic Clinical Prediction May Increase Health Disparities*, 3(99) NPJ Digit. Med. 2 (2020), <https://doi.org/10.1038/s41746-020-0304-9>. Reproduced with permission.

A multitude of fairness criteria exist with likely mutual incompatibility, so satisfying all of them becomes impossible because of inevitable unfairness by at least one criteria, thereby suggesting the impossibility of a fair and unbiased decision because there will be at least one unequal outcome.

To satisfy [a] more stringent, narrow, and rigorous definition of unfairness, it is not enough to observe differences in outcomes—one must understand the causes for these outcome differences. Such a causal concept of fairness is closely aligned to the legal concept of disparate treatment. According to causal definitions of fairness, similar individuals should not be treated differently due to having certain protected attributes that qualify for special

244. Paulus & Kent, *supra* note 240, at 2. D+ = disease present, D- = disease absent, T+ = test positive, T- = test negative.

protection from discrimination, such as a certain race/ethnicity or gender. However, causality is fundamentally unidentifiable in observational data, except with unverifiable assumptions. Thus, we are more typically stuck with deeply imperfect but ascertainable criteria serving as (often poor) proxies for causal fairness.²⁴⁵

Given the emerging healthcare applications of machine artificial-intelligence research despite the lack of consensus on application and evaluation, it is important to recognize that current literature provides some frameworks for guidance about the evaluation of artificial intelligence in medicine as well as stages of development, reporting, evaluation, implementation, and surveillance.²⁴⁶

Informed Consent

Principles

Medical informed consent is an ethical, moral, and legal responsibility of physicians.²⁴⁷ It is guided by four ethical principles: autonomy, beneficence, non-maleficence, and justice.²⁴⁸ Autonomy refers to informed, rational decision-making after “unbiased and thoughtful deliberation.”²⁴⁹ Beneficence represents the “moral obligation of physicians to act for the benefit of patients.”²⁵⁰ These two principles place physicians in conflict because they wish to provide the care they believe is best for the patient, but because that care usually involves some risk or cost, physicians also recognize that patient autonomy and preferences may affect their recommendation.²⁵¹ In a study examining the incidence of erectile dysfunction with use of a beta-blocker medication known to be beneficial, heart disease patients were (1) blinded to the drug, (2) informed of the drug name only, or (3) informed about its erectile dysfunction adverse effect. Among those blinded, 3.1% developed erectile dysfunction compared with 15.6% of those

245. *Id.* at 3.

246. Norah L. Crossnohere et al., *Guidelines for Artificial Intelligence in Medicine: Literature Review and Content Analysis of Frameworks*, 24 J. Med. Internet Rsch. e36823 (2022), <https://doi.org/10.2196/36823>; Sebastian Vollmer et al., *Machine Learning and Artificial Intelligence Research for Patient Benefit: 20 Critical Questions on Transparency, Replicability, Ethics, and Effectiveness*, 368 B.M.J. 16927 (2020), <https://doi.org/10.1136/bmj.16927>; Lazaros Belbasis & Orestis A. Panagiotou, *Reproducibility of Prediction Models in Health Services Research*, 15 BMC Rsch. Notes 204 (2022), <https://doi.org/10.1186/s13104-022-06082-4>.

247. Timothy J. Paterick et al., *Medical Informed Consent: General Considerations for Physicians*, 83 Mayo Clinic Proc. 313, 313 (2008), <https://doi.org/10.4065/83.3.313>.

248. Jaime S. King & Benjamin W. Moulton, *Rethinking Informed Consent: The Case for Shared Decision-Making*, 32 Am. J. L. & Med. 429, 435 (2006), <https://doi.org/10.1177/009885880603200401>.

249. *Id.* at 435.

250. *Id.*

251. *Id.* at 436.

given the drug name and 31.2% of those informed about adverse effects, showing that being informed increased the risk for adverse effects and might deprive patients of benefit from a drug because they stop taking it.²⁵² Physicians must balance the desire to provide beneficial care with the obligation to promote autonomous decisions by informing patients of potential adverse effects or tradeoffs.

Standards

State jurisdictions differ in their standards for disclosure, with half adopting the physician or professional standard (the information that other local physicians with similar skill levels would provide) and the other half adopting the patient or materiality standard (the information that a reasonable patient would deem important in decision-making).²⁵³ The informed-consent process involves the disclosure of alternative treatment options, including no treatment, and the risks and benefits associated with each alternative. Discussion should include severe risks and frequent risks, but the courts have not provided explicit guidance about what constitutes sufficient severity or frequency. Patients should be considered by the court to be competent and should have the capacity to make decisions (understanding choices, risks, and benefits). The decision should be voluntary—of free mind and free will, without coercion or manipulation. The language used should be understandable to the patient, and treatment should not proceed unless the physician believes the patient understands the options and their risks and benefits.

Patients may withdraw consent or refuse treatment. Such an action should engender additional discussion, and documentation may include the completion of a withdrawal-of-consent form. In certain situations, exceptions to medical consent may arise in emergencies, when the treatment is recognized by prudent physicians to involve no material risk to patients and when the procedure is unanticipated and not known to be necessary at the time of consent.²⁵⁴

In 2024, the American Law Institute recognized that patients have choices among alternative treatments, so plaintiffs no longer need to prove that patients in similar situations would not have chosen treatment if they had received complete disclosure of risks, and now only need to prove that they would have chosen an alternative treatment that would have been reasonable for them.²⁵⁵

252. Antonello Silvestri et al., *Report of Erectile Dysfunction After Therapy with Beta-Blockers Is Related to Patient Knowledge of Side Effects and Is Reversed by Placebo*, 24 Eur. Heart J. 1928, 1928 (2003), <https://doi.org/10.1016/j.ehj.2003.08.016>.

253. King & Moulton, *supra* note 248, at 430.

254. Paterick et al., *supra* note 247, at 318.

255. Daniel G. Aaron et al., *A New Legal Standard for Medical Malpractice*, 333 JAMA 1161–65 (Feb. 26, 2025), <https://doi.org/10.1001/jama.2025.0097>.

Patient care

The *Merenstein* case described an unpublished trial in which, during his residency, Dr. Merenstein examined a highly educated man.²⁵⁶ The examination included a discussion of the relevant risks and benefits regarding prostate cancer screening using the prostate-specific antigen (PSA) test based on recommendations from the U.S. Preventive Services Task Force, the American College of Physicians–American Society of Internal Medicine, the American Medical Association, the American Urological Association, the American Cancer Association, and the American Academy of Family Physicians. Dr. Merenstein testified that the patient declined the test because of the high false-positive rate, the risk of treatment-related adverse effects, and the low risk of dying from prostate cancer. Another physician seeing the same patient subsequently ordered a PSA without any patient discussion. The PSA was high, and the patient was diagnosed with incurable advanced prostate cancer. The plaintiff’s attorney argued that despite the guidelines above, the standard of care in Virginia was to order the blood test without discussion, based on four physician witnesses who supported the plaintiff’s attorney “that the standard of care in Virginia was to order the test without discussing it with the patient,” so the jury ruled in favor of the plaintiff.²⁵⁷ It is important to note that the standard of care varies from state to state with about half having physician-based standards requiring “physicians to inform a patient of the risks, benefits and alternatives to a treatment in the same manner that a ‘reasonably prudent practitioner’ in the field would,” and the other half having patient-based standards requiring physicians to provide “all information on the risks, benefits and alternatives to a treatment that a ‘reasonable patient’ would attach significance to in making a treatment decision.”²⁵⁸

In 2024, the American Law Institute shifted its historic focus on local habitual practice to more patient-centered reasonable care, defined as that from competent and qualified clinicians that could override customary community practices, e.g., through evidence-based medicine.²⁵⁹

Patient preferences

To illustrate the importance of patient preferences, a woman with breast cancer described her experience:

256. Daniel Merenstein, *A Piece of My Mind: Winners and Losers*, 291 JAMA 15, 15–16 (2004), <https://doi.org/10.1001/jama.291.1.15>.

257. King & Moulton, *supra* note 248, at 434.

258. *Id.* at 432–34.

259. Aaron et al., *supra* note 255.

[A]s the surgeon diagramed the incision points on my chest with a felt-tip pen, my husband asked a question: Is it really necessary to transfer this back muscle? The doctor's answer shocked us. No, he said, he could simply operate on my chest. That would cut surgery and recovery time in half. He had planned the more complicated procedure because he thought it would have the best cosmetic result. "I assumed that's what you wanted."²⁶⁰

Instead the woman preferred the less invasive approach that shortened her recovery time.

Randomized trials

In the research setting, a randomized trial with and without informed consent demonstrated that the process of getting informed consent altered the effect of a placebo when given to patients with insomnia. The first patient of each pair was randomized to no informed consent and the second to informed consent. Out of 56 patients randomized to informed consent, 26 declined to participate in the study (the patients without informed consent had no choice and were unaware of their participation in a study). The informed consent process created a "biased" group because the age and gender for those who declined participation differed significantly from those who did agree to be included in the study. The hypnotic activity of placebo was significantly higher without informed consent, and adverse events were found more commonly in the group receiving informed consent. The study suggests that the process of getting informed consent introduced biases in the patient population and affected the efficacy and adverse effects observed in this clinical trial, thereby potentially affecting the general applicability of any findings involving informed consent.²⁶¹

Health information sources

Besides physicians, patients may get health information from the internet, family, friends, and the media (newspapers, magazines, television). Of internet users, 80% had searched for information on at least 1 of 15 major health topics, but use varied from 62% to 89% by age, sex, education, and race/ethnicity groups.²⁶² The

260. Julie Halpert, *What Do Patients Want?* 141 *Newsweek*, no. 17, Apr. 27, 2003, at 63–64, <https://perma.cc/8AZE-Q3HE>.

261. R. Dahan et al., *Does Informed Consent Influence Therapeutic Outcome? A Clinical Trial of the Hypnotic Activity of Placebo in Patients Admitted to Hospital*, 293 *Brit. Med. J. (Clinical Rsch. Ed.)* 363, 363 (1986), <https://doi.org/10.1136/bmj.293.6543.363>.

262. Susannah Fox, Pew Rsch. Ctr., *Health Topics* 13–14 (2011), <https://perma.cc/F4QU-45FB>.

percent of all adults who look online for health information varied from 72% for those 18–29 to 30% for those 65 and older.²⁶³ Over the past decade, smartphone use has grown dramatically across all age groups, but especially those 65 and older (13% in 2012 to 61% in 2021).²⁶⁴ A cross-sectional November 2006 and May 2007 national survey of U.S. adults who had made a medical decision found that internet use averaged 28%, but varied from 17% for breast cancer screening to 48% for hip/knee replacement, among those 40 years of age and older.²⁶⁵ However, even among internet users, healthcare providers were felt to be the most influential source of information for medical decisions, followed by the internet, family and friends, and then media. In this digital age, clinicians remain the most influential information source for now.

Risk Communication

Multiple health outcomes may result from alternative treatment choices, and how patients feel about the relative importance of those outcomes varies.²⁶⁶ When patients with recently diagnosed, curable prostate cancer were presented with 93 possible questions that might be important to patients like themselves, 91 of the questions were cited as relevant to at least one patient, demonstrating that personal health goals and concerns vary substantially.²⁶⁷ Communication skills should include patient problem assessment (appropriate questioning techniques, seeking patient's beliefs, checking patient's understanding of the problem); patient education and counseling (eliciting patient's perspective, providing clear instructions and explanations, assessing understanding); negotiation and shared decision-making (surveying problems and delineating options, arriving at mutually acceptable solutions); and relationship development and maintenance (encouraging patient expression, communicating a supportive attitude, explaining any jargon, and using nonverbal behavior to enhance communication).²⁶⁸

263. Mary Madden, Pew Rsch. Ctr., *Older Adults and Internet Use: (Some of) What We Know* 20 (2013), <https://perma.cc/5L49-4E7T>.

264. Michelle Faverio, Pew Rsch. Ctr., *Share of Those 65 and Older Who Are Tech Users Has Grown in the Past Decade* 1 (2022), <https://perma.cc/TT4J-AFLE>.

265. Mick P. Couper et al., *Use of the Internet and Ratings of Information Sources for Medical Decisions: Results from the DECISIONS Survey*, 30 *Med. Decision Making* 106S, 106S (2010), <https://doi.org/10.1177/0272989X10377661>.

266. Kassirer, *supra* note 220, at 899.

267. Deb Feldman-Stewart et al., *What Questions Do Patients with Curable Prostate Cancer Want Answered?*, 20 *Med. Decision Making* 7, 7 (2000), <https://doi.org/10.1177/0272989X0002000102>.

268. Michael J. Yedidia et al., *Effect of Communications Training on Medical Student Performance*, 290 *JAMA* 1157, 1159 (2003), <https://doi.org/10.1001/jama.290.9.1157>.

Types of risk-communication estimates

Certain forms of risk communication, however, may be confusing and should be avoided: “single event probabilities, conditional probabilities (such as sensitivity and specificity), and relative risks.”²⁶⁹ Single-event probabilities are a “source of miscommunication” because the events remain unspecified, for example, a statement that a medication results in a 30% to 50% chance of developing erectile dysfunction could be misconstrued by patients to mean an erectile dysfunction problem in 30% to 50% of each their sexual encounters instead of a more specific statement such as out of 10 people like you taking this medication, three to five of them experience erectile dysfunction.²⁷⁰ This natural frequency statement specifies a reference class (e.g., out of 10 people and not each encounter), thereby reducing misunderstanding.²⁷¹

For communicating benefit by citing relative risk (ratio of the risk of dying in a group taking a medication divided by the risk of dying in a group not taking a medication), consider an advertising statement that taking a cholesterol-lowering medication reduces the risk of dying by 22%.²⁷² This may be misinterpreted as saying that out of 1,000 patients with high cholesterol, 220 of them can avoid dying by taking cholesterol-lowering medications. The actual data show that 32 deaths occur among 1,000 patients taking the medication, and 41 deaths occur among 1,000 patients taking the placebo, so the relative risk is 0.78 (32/41). The relative risk reduction equals one minus the relative risk or 0.22 (22%). A preferred way to express the benefit would be the absolute risk reduction (the difference between 41 and 32 deaths in 1,000 patients), or to say that in a group of 1,000 people like you with high cholesterol, taking a cholesterol medication for five years helps nine of them avoid dying.²⁷³

To illustrate potential misinterpretation of risk further, a relative risk reduction of 22% has very different absolute risk reductions depending on the event rates without treatment. If the mortality rate is 20% without treatment, then the absolute risk reduction is 4.4% or saving 44 of 1,000 treated patients (22% times 20%), but for an event rate of 2% without treatment, then the same 22% risk reduction results in an absolute reduction of 0.44% or saving 4.4 of 1,000 treated patients. The number needed to treat is an alternative form of risk communication to account for the risk without treatment. It is the reciprocal of the absolute risk difference or one divided

269. Gerd Gigerenzer & Adrian Edwards, *Simple Tools for Understanding Risks: From Innumeracy to Insight*, 327 B.M.J. 741, 741 (2003), <https://doi.org/10.1136/bmj.327.7417.741>.

270. *Id.*

271. *Id.* at 743; see also section titled “Multiple Diseases and Multiple Tests” above and David H. Kaye & Hal S. Stern, *Reference Guide on Statistics and Research Methods*, “Appendix: Conditional Probability and Bayes’ Rule,” in this manual.

272. Gerd Gigerenzer, *Calculated Risks: How to Know When Numbers Deceive You* 34 (2002).

273. *Id.* at 34–35.

by the quantity nine lives saved per 1,000 ($1 \div (9/1000)$) treated with cholesterol medications in the prior paragraph. Therefore 111 patients need to be treated with a cholesterol medication for five years to save one of them, or in the illustrative example above with a relative risk reduction of 22% and a mortality rate of 20% or 2%, either 23 or 227, respectively, would need to be treated to save one patient.

Breast-cancer screening communication

In the analysis of randomized trials of mammography for the U.S. Preventive Services Task Force, the number needed to screen (NNS) to avoid one breast-cancer death over 10 years was 3,448 (10,000/2.9) for 39- to 49-year-olds, 1,299 (10,000/7.7) for 50- to 59-year-olds, and 469 (10,000/21.3) for 60- to 69-year-olds.²⁷⁴ The numbers needed to harm (NNH) with annual or biennial mammography over 10 years were: screening every 1.6 (1/0.61) to 2.4 (1/0.42) 40-year-old women yields one false positive result and screening 14 (1/0.07) to 20 (1/0.05) 40-year-old women yields one breast biopsy.²⁷⁵ Screening for breast cancer and treating those found to have breast cancer saves lives. From epidemiologic population-based incidence and mortality data and from screening trials with long-term follow-up, it becomes clear that some patients experience over-diagnosis, i.e., some women have screen-detected cancer that would never have developed clinical signs or symptoms or caused their mortality in their lifetime. Estimates of overdiagnosis from trials ranged from 11% to 22%, and so, out of 100 women diagnosed with breast cancer from screening, five to nine of them undergo treatment for a cancer that may have never caused any mortality.²⁷⁶ Importantly, no one can tell if any particular woman has been overdiagnosed because this is unobservable.²⁷⁷

To summarize the evidence,

Mammography does save lives, more effectively among older women, but does cause some harm. Do the benefits justify the risks? The misplaced propaganda battle seems to now rest on the ratio of the risks of saving a life compared with the risk of overdiagnosis, two very low percentages that are imprecisely estimated and depend on age and length of follow-up.²⁷⁸

274. Heidi D. Nelson et al., *Effectiveness of Breast Cancer Screening: Systematic Review and Meta-Analysis to Update the 2009 U.S. Preventive Services Task Force Recommendation*, 164 *Annals Internal Med.* 244, 246 (2016), <https://doi.org/10.7326/M15-0969>.

275. Heidi D. Nelson et al., *Harms of Breast Cancer Screening: Systematic Review to Update the 2009 U.S. Preventive Services Task Force Recommendation*, 164 *Annals Internal Med.* 256, 257–59 (2016), <https://doi.org/10.7326/M15-0970>.

276. *Id.* at 256.

277. Klim McPherson, *Screening for Breast Cancer: Balancing the Debate*, 340 *B.M.J.* c3106 at 234 (2010), <https://doi.org/10.1136/bmj.c3106>.

278. *Id.*

Technologies

Stephen Hawking said, “Our future is a race between the growing power of technology and the wisdom with which we use it.”²⁷⁹ Rapidly growing sources of data fuel technology growth, particularly artificial intelligence (AI) with the potential to improve diagnosis, clinical care, health research, drug development, and public health (surveillance, outbreak response, and health system monitoring).²⁸⁰ However, the World Health Organization (WHO) identifies six key ethical principles for the use of AI in health:

- Protecting human autonomy: Providers must have the necessary information to make safe and effective use of AI systems. Patients must understand what part those systems play in their care. The AI systems must protect privacy and confidentiality with informed consent and appropriate legal protection of data;²⁸¹
- Promoting human well-being and safety and the public interest: The AI systems should meet regulatory requirements and be safe, accurate and efficient. The systems should practice quality control and quality improvement and, importantly, do no harm;²⁸²
- Ensuring transparency, explainability, and intelligibility: The design or deployment of an AI technology should be published or documented with sufficient information to facilitate public deliberation and consultation;²⁸³
- Fostering responsibility and accountability: Establishing predefined points of human supervision through regulatory principles upstream and downstream of any AI algorithm would provide a “human warranty”;²⁸⁴
- Ensuring inclusiveness and equity: Design features should promote the “widest possible appropriate, equitable use and access, irrespective of age, sex, gender, income, race, ethnicity, sexual orientation, ability or other characteristics protected under human rights codes” and should seek to “minimize inevitable disparities in power that arise between providers and patients, between policy-makers and people, and between companies and governments”;²⁸⁵ and
- Promoting AI that is responsive and sustainable: Designers, developers, and users should “continuously, systematically and transparently assess AI

279. Stephen Hawking, *Brief Answers to the Big Questions* 196 (2018).

280. Health Ethics & Governance (HEG), *Ethics and Governance of Artificial Intelligence for Health: WHO Guidance*, Executive Summary iii (World Health Organization ed., 2021), <https://www.who.int/publications/i/item/9789240037403>.

281. *Id.* at iv.

282. *Id.* at v.

283. *Id.*

284. *Id.*

285. *Id.* at v–vi.

applications during actual use” and governments and companies should address anticipated disruptions in the workplace to ensure sustainability.²⁸⁶

Data collection and use

Personal health data has grown massively to include “genomic data, radiological images, medical records^[287] and nonhealth data converted into health data . . . from standard sources (e.g. health services, public health, research) and further sources (environmental, lifestyle, socioeconomic, behavioural and social).” Concerns include (1) the quality of the data and inherent biases in the training data (e.g., underrepresentation and their effect on algorithms); (2) safeguarding individual privacy (e.g., discrimination due to health status; cyber theft and accidental or intentional disclosure); (3) collected health data may exceed what is required—of particular concern is a “behavioral data surplus” that is “repurposed for uses that raise serious ethical, legal and human rights” questions;²⁸⁸ and (4) data colonialism “may foster a divide between those who accumulate, acquire, analyze and control such data and those who provide the data but have little control over their use.” As noted by the WHO, “true informed consent is increasingly infeasible in an era of biomedical big data, especially in an environment driven mainly by companies seeking to generate profits from the use of data” because of the “scale and complexity” of such data and because “all of the potential uses may not be known,” such as for “population-level data analytics or predictive-risk modeling.”²⁸⁹

286. *Id.* at vi.

287. Malpractice litigation includes the prominent use of electronic medical records with metadata (often referred to as “audit trails”) that allow someone to determine exactly when an entry was made into a record, what work station it was entered from, and whether any edits were made and when. An area of uncertainty remains as to the scope of a patient’s official or “legal” medical record under HIPAA and whether it includes text messages between clinicians, e.g., does the hospital have a duty to preserve such data and disclose it to the patient?

288. Ethics and Governance of Artificial Intelligence for Health: WHO Guidance, *supra* note 280, at 37. See, e.g., *Dinerstein v. Google, LLC*, 484 F. Supp. 3d 561 (2020), in which the University of Chicago was accused of sharing identifiable patient data with Google for the purposes of developing medical diagnostic tools. The health records shared with Google “were de-identified, except that dates of service were maintained” in the dataset, and the dataset also included “de-identified, free-text medical notes.” Marcelo Corrales Compagnucci et al., Box 3: *Dinerstein v. Google*, in Ethics and Governance of Artificial Intelligence for Health: WHO Guidance, *supra* note 280, at 38 (citing Alvin Rajkomar et al., *Scalable and Accurate Deep Learning with Electronic Health Records*, 1(18) NPJ Digit. Med. 6 (2018), <https://doi.org/10.1038/s41746-018-0029-1>). Another example is hospitals that gain consent to retain discarded biological samples from patients for research purposes when consenting to treatment at a facility.

289. Ethics and Governance of Artificial Intelligence for Health: WHO Guidance, *supra* note 280, at 40.

Shared Decision-Making

The “professional values of competence, expertise, empathy, honesty, and commitment are all relevant to communicating risk: Getting the facts right and conveying them in an understandable way are not enough.”²⁹⁰ Shared and informed decision-making has emerged as one part of patient care. It distinguishes problem solving that identifies one “right” course that leaves little room for patient involvement from decision-making in which several courses of action may be reasonable and in which patient involvement should determine the optimal choice. In such cases, healthcare choices depend not only on the likelihood of alternative outcomes resulting from each strategy but also on the patient preferences for possible outcomes and their attitudes about risk taking to improve future survival or quality of life and the timing of that risk (whether the risk occurs now or in the future).²⁹¹

Informed decision-making is

when an individual understands the nature of the disease or condition being addressed; understands the clinical service and its likely consequences, including risks, limitations, benefits, alternatives, and uncertainties; has considered his or her preferences as appropriate; has participated in decision making at a personally desirable level; and either makes a decision consistent with his or her preferences and values or elects to defer a decision to a later time.²⁹²

Shared decision-making is “when a patient and his or her healthcare provider(s), in the clinical setting, both express preferences and participate in making treatment decisions.”²⁹³ To assist with shared decision-making, health decision aids have been developed to help patients and their physicians choose among reasonable clinical options together by describing the “benefits, harms, probabilities, and scientific uncertainties.”²⁹⁴ In 2007, the legislature in the state of Washington became the first to establish and recognize in law a role for shared decision-making in informed consent.²⁹⁵ The bill goes on to encourage the development, certification, use, and evaluation of decision aids. The consent form

290. Adrian Edwards, *Communicating Risks*, 327 B.M.J. 691, 691 (2003), <https://doi.org/10.1136/bmj.327.7417.691>.

291. Michael J. Barry, *Health Decision Aids to Facilitate Shared Decision Making in Office Practice*, 136 *Annals Internal Med.* 127, 127 (2002), <https://doi.org/10.7326/0003-4819-136-2-200201150-00010>.

292. Peter Briss et al., *Promoting Informed Decisions About Cancer Screening in Communities and Healthcare Systems*, 26 *Am. J. Preventive Med.* 67, 68 (2004), <https://doi.org/10.1016/j.amepre.2003.09.012>.

293. *Id.*

294. Annette M. O'Connor et al., *Risk Communication in Practice: The Contribution of Decision Aids*, 327 B.M.J. 736, 736 (2003), <https://doi.org/10.1136/bmj.327.7417.736>.

295. Bridget M. Kuehn, *States Explore Shared Decision Making*, 301 *JAMA* 2539, 2539 (2009), <https://doi.org/10.1001/jama.2009.867>.

provides written documentation that the consent process occurred, but the crux of the medical consent process is the discussion that occurs between a physician and a patient. Physicians share their medical knowledge and expertise, and patients share their values (health goals) and preferences. It is an opportunity to strengthen the patient–physician relationship through shared decision-making, respect, and trust. Incorporating patient health goals into shared decision-making is supported by the core principles of medical decision analysis: the concept of utility-maximizing decisions, in which the probabilities of possible health outcomes from alternative choices is multiplied by the patient’s utility (a quantitative measure of preference) for those outcomes and summed over all possible probability–outcome pairs for each treatment choice (including no treatment). The intervention with the highest expected utility should be the preferred option.

In six mock trials of malpractice alleging failure to perform a prostate-specific antigen in response to the *Merenstein* case, a study found that 47 jurors were not willing to accept verbal testimony that a discussion occurred, but when documentation occurred, most (72%) felt that the defendant fulfilled the standard of care—although 28% still felt that screening should be done even if documentation had occurred. When a decision aid was included, 94% felt that the standard of care had been met.²⁹⁶ With regard to variation in care, a decision aid for surgery for a benign enlargement of the prostate decreased the likelihood of surgery versus control in one study but increased it in the other, the difference being the extent of surgery in the control group. In the United States, surgery decreased from 13% in the control group to 8% with a decision aid, and in the United Kingdom, surgery increased from 2% in the control group to 11% with a decision aid.²⁹⁷

Summary and Future Directions

Having sequenced the human genome, medical research is poised for exponential growth as the code for human biology (genomics) is translated into proteins (proteomics) and chemicals (metabolomics) to identify molecular pathways that lead to disease or that promote health. With advances in medical technologies in diagnosis and preventive and symptomatic treatment, the practice of medicine will be profoundly altered and redefined. For example, consider lymphoma, a blood cancer that used to be classified simply by appearance under the microscope as either

296. Michael J. Barry et al., *Reactions of Potential Jurors to a Hypothetical Malpractice Suit: Alleging Failure to Perform a Prostate-Specific Antigen Test*, 36 J. L. Med. Ethics 396, 400 (2008), <https://doi.org/10.1111/j.1748-720X.2008.00283.x>.

297. Michael J. Barry et al., *Randomized Trial of a Multimedia Shared Decision-Making Program for Men Facing a Treatment Decision for Benign Prostatic Hyperplasia*, 1 Disease Mgmt. Clinical Outcomes 5 (1997); Elizabeth Murray et al., *Randomised Controlled Trial of an Interactive Multimedia Decision Aid on Benign Prostatic Hypertrophy in Primary Care*, 323 B.M.J. 493, 497 (2001), <https://doi.org/10.1136/bmj.323.7311.493>.

Hodgkin's or non-Hodgkin's lymphoma. As science has evolved, it is now further classified by cellular markers that identify the underlying cancer cells as one of two cells that help with immunity (protecting the body from infection and cancer): T cells or B cells. Current research has characterized those cells further by identifying underlying genetic and cellular markers and pathways that distinguish these lymphomas and provide potential therapeutic targets. The growth in the research enterprise, both basic science and clinical translational—the translation of bench research to the bedside or the identification of a clinical need leading to basic science research to further elucidate molecular pathways and develop novel treatments or diagnostics—has greatly expanded research capacity to generate scientific medical research of all types.

With greatly expanded knowledge, research, and specialization, judgments about admissibility and about what constitutes expertise become increasingly difficult and complex. The sifting of this research into sufficiently substantiated, competent, and reliable evidence, however, relies on the traditional scientific foundation: first, biological plausibility and prior evidence; and second, consistent repeated findings. The practice of medicine at its core will continue to be a physician and patient interaction with professional judgment and communication as central elements of the relationship. Judgment is essential because of uncertainties in the underlying professional knowledge or because even if the evidence is credible and substantiated, there may be tradeoffs in risks of harms and benefits for testing and for treatment. Communication is critical because most decisions involve tradeoffs, in which case individual patient preferences for the outcomes that may be unique to a patient and that may affect decision-making should be considered.

In summary, medical terms shared by the legal and medical professions have differing meanings, for example, differential diagnosis, differential etiology, and general and specific causation. The basic concepts of diagnostic reasoning and clinical decision-making and the types of evidence used to make judgments as treating physicians or experts involve the same overarching theoretical issues: (1) alternative reasoning processes; (2) weighing risks, benefits, and evidence; and (3) communicating those risks.

Glossary of Terms

adequacy. In diagnostic verification, testing a hypothesized diagnosis for its adequacy in explaining the course of the disease.

atrial fibrillation. An abnormal heart rhythm where the heart muscles do not contract together so that blood swirls, leading to blood clots usually in the upper left part of the heart (left atrium). Those clots can then leave the heart and cause strokes or other symptoms.

attending physician. The physician primarily responsible for the patient's care at the hospital in which the patient is being treated.

Bayes' theorem (rule). In medicine, it is a method to calculate a post-test probability after a test result in a patient with a pretest probability of disease (suspicion) with additional information such as from a test result by using test characteristics: sensitivity for how well it performs in individuals with disease and specificity for how well it performs in those without disease.

causal reasoning. For physicians, causal reasoning typically involves understanding how abnormalities in physiology, anatomy, genetics, or biochemistry led to the clinical manifestations of disease. Through such reasoning, physicians develop a "causal cascade" or "chain or web of causation" linking a sequence of plausible cause-and-effect mechanisms to arrive at the pathogenesis or pathophysiology of a disease.

chief complaint. The primary or main symptom that caused the patient to seek medical attention.

coherency. In diagnostic verification, testing a particular diagnosis for its coherency involves determining the consistency of that particular diagnosis with predisposing risk factors, physiological mechanisms, and resulting manifestations.

conditional probability. The probability or likelihood of something, given that something else occurred or is present, for example, the likelihood of disease if a test is positive (post-test probability) or the likelihood of a positive test if disease is present (sensitivity). See Bayes' theorem (rule).

consulting physician. A physician, usually a specialist, who is asked by the patient's attending physician to provide an opinion regarding diagnosis, testing, or treatment or to perform a procedure or intervention, for example, surgery.

deductive reasoning. The process of having a specific diagnosis in mind and identifying the symptoms, signs, and lab tests that may occur with that diagnosis with the conclusion being certain. See differential diagnosis.

diagnostic test. A test ordered to confirm or exclude possible causes of a patient's symptoms or signs (distinct from screening test).

diagnostic verification. The last stage of narrowing the differential diagnosis to a final diagnosis by testing the validity of the diagnosis for its coherency, adequacy, and parsimony.

differential diagnosis. A set of diseases that clinicians consider as possible causes of the patient's complaint until the diagnosis is nearly final (diagnostic verification).

differential etiology. Term used by the court or witnesses to establish or refute external causation for a plaintiff's condition. For physicians, etiology refers to cause.

external causation. External causation is established by demonstrating that the cause of harm or disease originates from outside the plaintiff's body, for example, a defendant's action or product.

general causation. General causation is established by demonstrating, usually through scientific evidence, that a defendant's action or product causes (or is capable of causing) disease.

heuristics. Quick automatic rules of thumb or cognitive shortcuts often involving pattern recognition that facilitate rapid diagnostic and treatment decision-making. Use of heuristics or "thinking fast" may predispose to known cognitive errors, even among experts. See hypothetico-deductive.

hypothesis generation. A limited list of potential diagnostic hypotheses in response to symptoms, signs, and lab test results. See differential diagnosis.

hypothesis modification. A change in the list of diagnostic hypotheses (differential diagnosis) in response to additional information, for example, symptoms, signs, and lab test results. See differential diagnosis.

hypothesis refinement. A change in the likelihood of the potential diagnostic hypotheses (differential diagnosis) in response to additional information, for example, symptoms, signs, and lab test results. As additional information emerges, physicians evaluate those data for their consistency with the possibilities on the list and whether those data would increase or decrease the likelihood of each possibility. See differential diagnosis.

hypothetico-deductive. Deliberative and analytical reasoning involving hypothesis generation, hypothesis modification, hypothesis refinement, and diagnostic verification. Thinking "slow" is typically applied for problems outside an individual's expertise or difficult problems with atypical issues, as it may avoid known cognitive errors. See heuristics.

individual causation. See specific causation.

inductive reasoning. The process of arriving at a diagnosis based on symptoms, signs, and lab tests with the conclusion being probable rather than certain. See differential diagnosis.

intracranial hemorrhage. Bleeding that occurs within the substance of the brain.

myocardial infarction. The technical term for a heart attack.

osteopaths. Doctors of osteopathy (D.O.s), who receive similar training, licensure, and credentialing as medical doctors (M.D.s).

overdiagnosis. Screening and diagnostic testing can lead to the detection of apparent disease that may never cause harm, for example, the identification of slow-growing cancers that, even if untreated, would never cause symptoms or reduce survival because the screening test cannot distinguish the cancerous-appearing cells that would become symptomatic or cause mortality from those that would never do so. See overtreatment.

overtreatment. The treatment of patients with pseudo disease whose disease would never cause symptoms or reduce survival. The treatment may place patients at risk for treatment-related morbidity and possibly mortality without any potential for benefit. See overdiagnosis.

parsimony. In diagnostic verification, testing a particular diagnosis for its parsimony involves choosing the simplest single explanation as opposed to requiring the simultaneous occurrence of two diseases to explain the findings.

pathogenesis. See causal reasoning.

pathology test. Microscopic examination of body tissue typically obtained by a biopsy or during surgery to determine if the tissue appears to be abnormal (different than would be expected for the source of the tissue). The visual components of the abnormality are typically described (e.g., types of cells, appearance of cells, scarring, effect of stains or molecular markers that help facilitate identification of the components) and on the basis of visual pattern, the abnormality may be classified, for example, malignancy (cancer) or dysplasia (precancerous).

post-test probability. The suspicion or probability of a disease after additional information (such as from a test) has been obtained. The predictive value positive (or positive predictive value) is the probability of disease in those known to have a positive test result. The predictive value negative (or negative predictive value) is the probability of disease in those known to have a negative test result.

predictive value. The predictive value of a test when applied to a specific population of patients with a specific prevalence of disease.

pretest probability. The suspicion or probability of a disease before additional information (such as from a test) is obtained. Also referred to as prior probability.

prior probability. See pretest probability.

sarcoidosis. A collection of inflammatory cells called granulomas causing scar tissue that collect most often in the lungs or skin but can involve multiple other parts of the body.

screening test. A test performed in the absence of symptoms or signs to detect disease earlier—for example, cancer screening (distinct from diagnostic test).

sensitivity. Likelihood of a positive finding (usually referring to a test result but could also be a symptom or a sign) among individuals known to have a disease (distinct from specificity).

sign. An abnormal physical finding identified at the time of physical examination (distinct from symptoms).

specific causation. Established by demonstrating that a defendant's action or product is the cause of a particular plaintiff's disease. Also referred to as individual causation.

specificity. Likelihood of a negative finding (usually referring to a test result but could also be a symptom or a sign) among individuals who do not have a particular disease (distinct from sensitivity).

symptoms. The patient's description of a change in function, sensation, or appearance (distinct from sign).

syndrome. A collection of symptoms and signs that together characterize a specific disease or a specific group of diseases.

References on Medical Testimony

- Lynn S. Bickley et al., *Bates' Guide to Physical Examination and History Taking* (13th ed. 2020).
- Gerd Gigerenzer, *Calculated Risks: How to Know When Numbers Deceive You* (2002).
- Trisha M. Greenhalgh & Paul Dijkstra, *How to Read a Paper: The Basics of Evidence-Based Healthcare* (7th ed. 2024).
- Gordon Guyatt et al., *Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice* (3d ed. 2014).
- Miguel A. Hernan & James M. Robins, *Causal Inference: What If* (2024).
- Guidon W. Imbeds & Donald B. Rubin, *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction* (2015).
- Jerome P. Kassirer et al., *Learning Clinical Reasoning* (2d ed. 2010).
- National Academies of Sciences, Engineering, and Medicine, *Improving Diagnosis in Health Care* (2015).
- Harold C. Sox et al., *Medical Decision Making* (3d ed. 2024).
- Sharon E. Straus et al., *Evidence-Based Medicine: How to Practice and Teach EBM* (5th ed. 2018).

Reference Guide on Neuroscience

HENRY T. GREELY AND NITA A. FARAHANY

Henry T. Greely, J.D., is the Deane F. and Kate Edelman Johnson Professor of Law; Professor, by courtesy, of Genetics; and the Director of Center for Law and the Biosciences, at Stanford University.

Nita A. Farahany, J.D., Ph.D., is the Robinson O. Everett Professor of Law and Philosophy and Founding Director for the Initiative for Science and Society at Duke University.

CONTENTS

Introduction,	1187
The Human Brain,	1188
Cells,	1189
Brain Structure,	1193
Some Aspects of How the Brain Works,	1197
Some Common Neuroscience Techniques,	1200
Neuroimaging,	1202
CAT Scans,	1202
PET Scans and SPECT Scans,	1204
MRI—Structural and Functional,	1206
DTI,	1213
fNIRS,	1213
EEG and MEG,	1214
Other Techniques,	1215
Lesion Studies,	1216
TMS,	1216
DBS,	1217
Implanted Microelectrode Arrays,	1218
Issues in Interpreting Study Results,	1219
Replication,	1220
Problems in Experimental Design,	1220
The Number and Diversity of Participants,	1222
Applying Group Averages to Individuals,	1223
Technical Accuracy and Robustness of Imaging Results,	1224
Statistical Issues,	1225
Possible Countermeasures,	1226
Questions About the Admissibility and the Creation of Neuroscience Evidence,	1228
Evidentiary Rules,	1228
Relevance,	1228

Rule 702 and the Admissibility of Scientific Evidence, 1229
Rule 403, 1232
Other Potentially Relevant Evidentiary Issues, 1233
Constitutional and Other Substantive Rules 1235
Possible Rights Against Neuroscience Evidence, 1235
The Fifth Amendment privilege against self-incrimination, 1235
Other possible legal protections against compulsory neuroscience procedures, 1237
Other substantive rights against neuroscience evidence, 1238
Neuroscience evidence and the sixth and seventh amendment rights to trial by jury, 1239
Possible rights to the creation or use of neuroscience evidence, 1240
The Eighth Amendment right to present evidence of mitigating circumstances in capital cases, 1240
The Sixth Amendment right to present a defense, 1241
The Fourth Amendment, 1241
Examples of the Possible Uses of Neuroscience in the Courts, 1242
Criminal Responsibility, 1247
Lie Detection, 1250
In General, 1250
fMRI-Based Lie Detection, 1252
Issues Involved in the Use of EEG-Based Brain “Recognition”, 1258
Detection of Pain, 1259
Addiction, 1263
Conclusion, 1265
Glossary of Terms, 1266

FIGURES

1. Schematic of the typical structure of a neuron, 1190
2. Synapse. Communication between neurons occurs at the synapse, 1191
3. Lateral (left) and mid-sagittal (right) views of the human brain, 1194
4. Lobes of a hemisphere. Each hemisphere of the brain consists of four lobes: the frontal, parietal, temporal, and occipital lobes, 1195
5. Brodmann areas, 1198
6. CAT scan depicting axial sections of the human brain, 1203
7. PET scan depicting an axial section of the human brain, 1206
8. MRI machine. Magnetic resonance imaging systems are used to acquire both structural and functional images of the brain, 1208
9. Brain MRI scan depicting an axial (upper), coronal (lower left), and sagittal (lower right) image of the human brain, 1209
10. fMRI image. Functional MRI data reveal regions associated with cognition and behavior. Here, regions of the frontal and parietal lobes that are more active when remembering past events relative to detecting novel stimuli are depicted, 1212

Introduction

Science's understanding of the human brain is increasing exponentially. We know almost infinitely more than we did thirty years ago; however, we know almost nothing compared with what we are likely to know thirty years from now. The development of machine learning algorithms and artificial intelligence (AI) has allowed us to pull much more information out of existing methods of observing and recording brain activity, including greatly boosting the value of the old technology of electroencephalography (EEG).¹ The electrodes and electrode arrays that serve as tools for recording and stimulating small numbers of neurons inside living brains have improved greatly.² The development of human neural organoids and the increase in human/nonhuman brain chimeras have given us new ways to watch how brain tissues develop and function.³ And, less excitingly but equally importantly, increased use has given us better understanding of the limitations of some of our existing tools.⁴

We still remain very far from a deep and broad understanding of how human brains work, but some advances in understanding the human brain and its attendant mental states have resulted in neuroscience evidence already appearing in courtrooms. If, as neuroscience indicates, our mental states correspond to, and result from, physical states of our brain, our increased ability to discern those physical states will have huge implications for the law. Since 2008, lawyers in steadily increasing numbers have been trying to introduce neuropsychological and neuroimaging evidence as relevant to questions of individual criminal responsibility, such as claims of insanity or diminished responsibility, on issues of either criminal liability or sentencing. In May 2010, parties in two cases sought to introduce neuroimaging in court as evidence of honesty; we have begun to see efforts to use it to prove that a person is in pain. These and other uses of neuroscience are almost certain to increase with our growing knowledge of the human brain as well as continued technological advances in accurately and precisely measuring the brain. This reference guide strives to give judges some background knowledge about neuroscience and the strengths and weaknesses of its

1. See, e.g., Meysam Golmohammadi et al., *Deep Learning Approaches for Automated Seizure Detection from Scalp Electroencephalograms*, in *Signal Processing in Medicine and Biology* 235 (Iyad Obeid, Ivan Selesnick & Joseph Picone eds., 2020), <https://doi.org/10.1007/978-3-030-36844-9>.

2. See Nita A. Farahany, *The Battle for Your Brain: Defending the Right to Think Freely in the Age of Neurotechnology* ch. 9 (2023).

3. See Nat'l Acad. of Scis., Eng'g, & Med., *The Emerging Field of Human Neural Organoids, Transplants, and Chimeras* (2021), <https://doi.org/10.17226/26078>.

4. See, e.g., Russell A. Poldrack et al., *Scanning the Horizon: Towards Transparent and Reproducible Neuroimaging Research*, 18 *Nature Revs. Neuroscience* 115 (2017), <https://doi.org/10.1038/nrn.2016.167>; Scott Marek et al., *Reproducible Brain-Wide Association Studies Require Thousands of Individuals*, 603 *Nature* 654 (2022), <https://doi.org/10.1038/s41586-022-04492-9>.

possible applications in litigation in order to help them become better prepared for these cases.⁵

Evidence about the brain has been used in the courtroom for decades, by psychologists, psychiatrists, neurologists, and other professionals. That evidence has been used to help courts decide issues such as the existence or extent of injuries, criminal defenses of insanity or reduced capacity, questions of competency in widely varying contexts, and more. Most, if not all, of the tools and techniques discussed in this reference guide could be used in court to provide evidence relevant to these efforts to answer more traditional legal questions about human brains and minds. We have chosen to focus this guide on some of the new ways in which—and different professionals through which—neuroscience evidence may enter the court, either for new questions, such as lie detection, or through new methods, like neuroimaging to provide evidence of *mens rea*, because judges and lawyers will be less familiar with them. This reference guide's descriptions of the techniques, and of their strengths and weaknesses, should apply even when more customary issues are presented.

This reference guide begins with a brief overview of the structure and function of the human brain. It then describes some of the tools neuroscientists use to understand the brain—tools likely to produce findings that parties will seek to introduce in court. Next, it discusses a number of fundamental issues that must be considered when interpreting neuroscientific findings. Finally, after discussing, in general, the evidentiary issues raised by neuroscience-based evidence, this guide concludes by analyzing a few illustrative situations in which neuroscientific evidence is likely to appear in court—now and, to an increasing extent, in the future.

The Human Brain

This abbreviated and simplified discussion of the human brain describes the cellular basis of the nervous system, the structure of the brain, and finally our current understandings of how the brain works. More detailed, but still accessible,

5. The *Reference Guide on Neuroscience* in the third edition of this manual, upon which this chapter is based, was published in 2011 with that aim. *A Primer on Criminal Law and Neuroscience* (Stephen J. Morse & Adina L. Roskies eds., 2013) is another useful source, commissioned by the Law and Neuroscience Project, funded by the John D. and Catherine T. MacArthur Foundation. That project had already published a pamphlet written by neuroscientists for judges, with brief discussions of issues relevant to law and neuroscience: *A Judge's Guide to Neuroscience: A Concise Introduction* (Michael S. Gazzaniga & Jed. S. Rakoff eds., 2010). More recently, a textbook on law and neuroscience has been published in a second edition: *Law and Neuroscience* (Owen D. Jones, Jeffrey D. Schall & Francis X. Shen eds., 2d ed. 2020). One early book on a broad range of issues in law and neuroscience also deserves mention: *Neuroscience and the Law: Brain, Mind, and the Scales of Justice* (Brent Garland ed., 2004).

information about the human brain can be found in academic textbooks and in popular books for general audiences.⁶

Cells

Like most of the human body, the nervous system is made up of cells. Adult humans contain somewhere between 30 trillion and 40 trillion human cells. Each of those cells is both individually alive and part of a larger living organism.

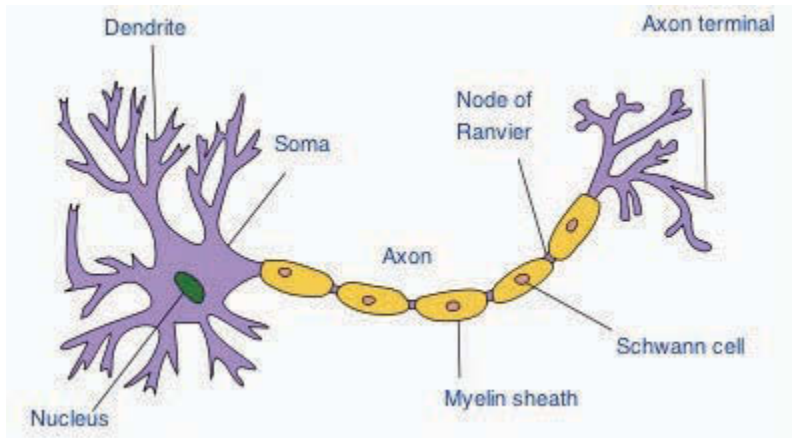
For all people, each human cell in the body (except for a few unusual cell types, like red blood cells) contains each person's entire complement of human genes—their genome. The genes, found on very long molecules of deoxyribonucleic acid (DNA) that make up a human's forty-six chromosomes, work by leading the cells to make other molecules, notably proteins and ribonucleic acid (RNA). Cells are different from each other not because they *contain* different genes, but because they turn on and off different sets of genes. All human cells seem to use the same group of several thousand “housekeeping” genes that run the cell's basic machinery, but skin cells, kidney cells, and brain cells differ in which other genes they use. Scientists count different numbers of types of human cells, with estimates ranging from a few hundred to a few thousand (depending largely on how narrowly or broadly one defines a cell type).

The most important cells in the nervous system are called neurons. Neurons pass messages from one to another in a complex way that appears to be responsible for brain function, conscious or otherwise.

Neurons come in many sizes, shapes, and subtypes (with their own names), but they generally have three features: a cell body, short extensions called dendrites, and a longer extension called an axon (see Figure 1). The cell body contains the nucleus of the cell, which in turn contains the forty-six chromosomes

6. The Society for Neuroscience, the very large scholarly society that covers a wide range of brain science, has published a brief and useful primer about the human brain called *Brain Facts*. The most recent edition, the eighth, published in 2018, is available free at <https://perma.cc/TEL6-6YNG>. Widely used academic textbooks include Eric R. Kandel et al., *Principles of Neural Science* (6th ed. 2021); Michael S. Gazzaniga, Richard B. Ivry & George R. Mangun, *Cognitive Neuroscience: The Biology of the Mind* (6th ed. 2025); and Larry Squire et al., *Fundamental Neuroscience* (4th ed. 2012). Some particularly interesting books about various aspects of the brain written for a popular audience include Oliver Sacks, *The Man Who Mistook His Wife for a Hat and Other Clinical Tales* (1985); Antonio R. Damasio, *Descartes' Error: Emotion, Reason, and the Human Brain* (1994); Daniel L. Schacter, *Searching for Memory: The Brain, the Mind, and the Past* (1996); Joseph E. LeDoux, *The Emotional Brain: The Mysterious Underpinnings of Emotional Life* (1996); Christopher D. Frith, *Making up the Mind: How the Brain Creates Our Mental World* (2007); Sandra Aamodt & Sam Wang, *Welcome to Your Brain: Why You Lose Your Car Keys but Never Forget How to Drive and Other Puzzles of Everyday Life* (2008). Matthew Cobb, *The Idea of the Brain: The Past and Future of Neuroscience* (2020), provides a good history of neuroscience.

Figure 1. Schematic of the typical structure of a neuron.



Source: Quasar Jarosz, https://en.m.wikipedia.org/wiki/File:Neuron_Hand-tuned.svg.

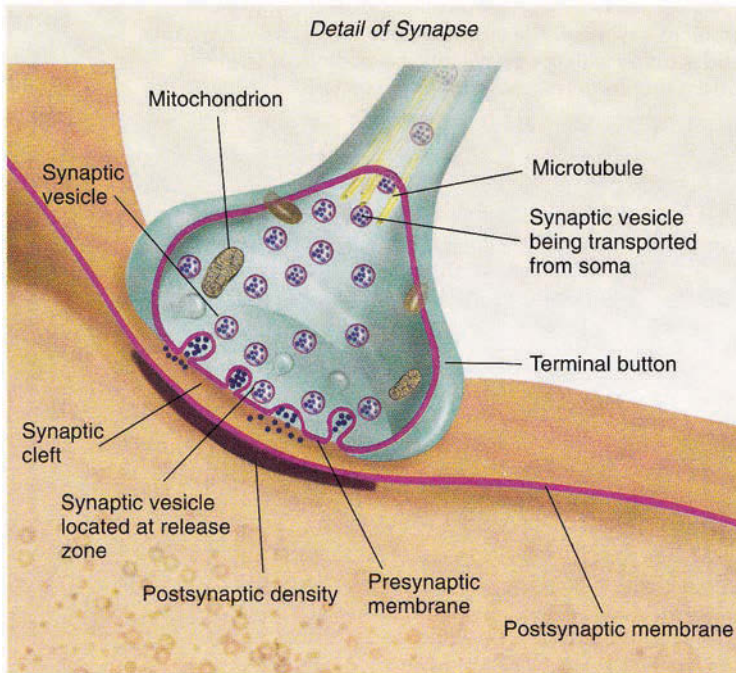
with the cell's DNA. The dendrites and axons both reach out to make connections with other neurons. The dendrites generally receive information from other neurons; the axons send information.

Communication between neurons occurs at areas called synapses (Figure 2), where two neurons almost meet. At a synapse, the two neurons will come within less than a micrometer (a millionth of a meter) of each other, with the presynaptic side, on the axon, separated from the postsynaptic side, on the dendrite, by a gap called the synaptic cleft. At synapses, when the axon (on the presynaptic side) “fires” (becomes active), it releases molecules, known as neurotransmitters, into the synaptic cleft. Some of those molecules are picked up by special receptors on the dendrite that is on the postsynaptic side of the cleft. More than 100 different neurotransmitters have been identified; among the best known are dopamine, serotonin, glutamate, and acetylcholine.

At the postsynaptic side of the cleft, neurotransmitters binding to the receptors can have a wide range of effects. Sometimes they cause the receiving (postsynaptic) neuron to fire, sometimes they suppress (inhibit) the postsynaptic neuron from firing, and sometimes they seem to do neither. The response of the receiving neuron is a complicated summation of the various messages it receives from multiple neurons that converge, through synapses, on its dendrites.

A neuron that does fire does so by generating an electrical current that flows down the length of its axon (away from the cell body). We normally think of electrical current as flowing in things like copper wiring. In that case, free

Figure 2. Synapse. Communication between neurons occurs at the synapse.



Source: From Carlson, Neil R. *Foundations of Physiological Psychology* (with Neuroscience Animations and Student Study Guide CD-ROM, 2005). Reprinted by permission of Pearson Education, Inc.

electrons move down the wire. The electrical currents of neurons are more complicated. Molecules with a positive or negative electrical charge (ions) move through the neuron's membrane and create differences in the electrical charge between the inside and outside of the neuron, with the current traveling along the axon, rather like a fire brigade passing buckets of water in only one direction down the line. Firing occurs in milliseconds. This process of moving ions in and out of the cell membrane requires that the cell use large amounts of energy. When the current reaches the end of the axon, it may or may not lead to the axon releasing neurotransmitters into the synaptic cleft. This complicated part-electrical, part-chemical system is how information passes from one neuron to another.

The axons of human neurons are all microscopically narrow, but they vary enormously in length. Some are micrometers long; others, such as the axons of neurons running from the base of the spinal cord to the toes, are several feet long.

Longer axons tend to be coated with a fatty substance called myelin. Myelin helps insulate the axon and thus increases the strength and efficiency of the electrical signal, much like the insulation wrapped around a copper wire. (The destruction of this myelin sheathing is the cause of multiple sclerosis.) Axons coated with myelin appear white; thus, areas of the nervous system that have many myelin-coated axons are referred to as “white matter.” Cell bodies, by contrast, look gray; areas with many cell bodies and relatively few axons make up our “gray matter.” White matter can roughly be thought of as the wiring that connects gray matter to the rest of the body or to other areas of gray matter.

What we call nerves are really bundles of neurons. For example, we all have nerves that run down our arms to our fingers. Some of those nerves consist of neurons that pass messages from the fingers, up the arm, to other neurons in the spinal cord that then pass the messages on to the brain, where they are analyzed and experienced. This is how we feel things with our fingers. Other nerves are bundles of neurons that pass messages from the brain through the spinal cord to nerves that run down the arms to the fingers, telling them when and how to move.

Neurons can connect with other neurons or with other kinds of cells. Neurons that control body movements ultimately connect to muscle cells—these are called motor neurons. Neurons that feed information into the brain start with specialized sensory cells (i.e., cells specialized for detecting different types of stimuli—light, touch, heat, pain, and more) that fire in response to the appropriate stimulus. Their firings ultimately lead, directly or through other neurons, into the brain. These are sensory neurons. These neurons send information only in one direction—motor neurons ultimately from the brain, sensory neurons to the brain. The paralysis caused by, for example, severe damage to the spinal cord both prevents the legs from receiving messages to move that would come from the brain through the motor neurons and keeps the brain from receiving messages from sensory neurons in the legs about what the legs are experiencing. The break in the spinal column prevents the messages from getting through, just as a break in a local telephone line will keep two parties from connecting. (There are, unfortunately, not yet any human equivalents to wireless service.)

Estimates of the number of cells in a human brain vary widely, from a few hundred billion to several trillion. These cells include those that make up blood vessels and various connective tissues in the brain, but most of them are specialized brain cells. About 80 to 100 billion of these brain cells are neurons; the other cells (and the source of most of the uncertainty about the number of cells) are principally another class of cells referred to generally as glial cells.

Glial cells play many important roles in the brain, including, for example, producing and maintaining the myelin sheaths that insulate axons and serving as a special immune system for the brain. Traditionally, glial cells were often dismissed as “mere support staff”—their name comes from the Greek word for “glue,” which indicates the early understanding of their role. But emerging

evidence suggests that they may play a much larger role in mental processes than previously believed. Work pioneered by the late Ben Barres has shown that glial cells not only are crucial for holding neurons together, but are essential for forming, editing, and pruning neuronal connections (synapses). They also produce factors that, themselves, affect neurotransmitter concentrations and the firings of neurons. The full importance of glial cells is still being discovered, and another edition of this manual may devote more space to them. At this point, however, glial cells have not been used significantly in the kinds of neuroscience evidence that has been working its way into courts. So we will concentrate on neurons, the brain structures they form, and how those structures work.

Brain Structure

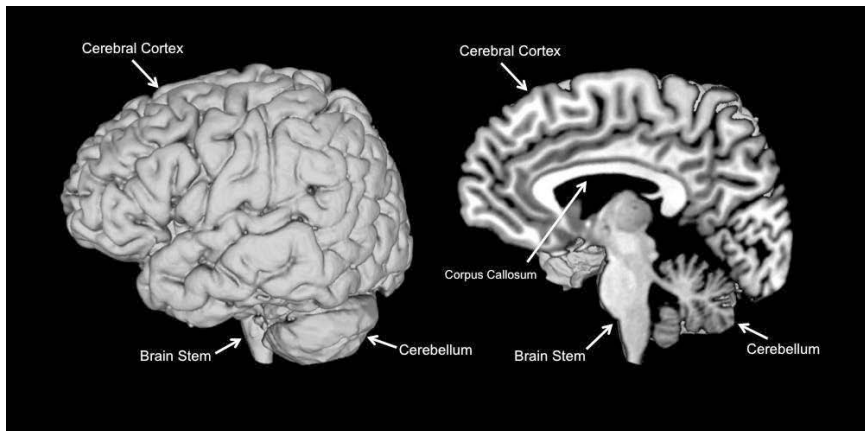
Anatomists refer to the brain, the spinal cord, and a few other nerves directly connecting to the brain as the central nervous system. All the other nerves are part of the peripheral nervous system. This reference guide will not discuss the peripheral nervous system in any detail, despite its importance in, for example, assessing some aspects of personal injuries. We will also, less fairly, ignore parts of the central nervous system other than the brain, even though the spinal cord, in particular, plays an important role not just in passing messages going into and coming out from the brain, but in changing their strength and meanings.

The average adult human brain weighs about three pounds and fills a volume of about 1,300 cubic centimeters. Two standard wine bottles hold 750 cubic centimeters each, so, if liquid, a human brain would not quite fill up a double bottle (a magnum); it would fill a measuring bowl to about 5.4 cups. Living brains have a consistency similar to that of Jell-O. Despite their softness, brains are made up of regular shapes and structures that are generally consistent from person to person. Just as every nondamaged or nondeformed human face has two eyes, two ears, one nose, and one mouth with standard numbers of various kinds of teeth, every normal brain has the same set of identifiable structures, both large and small. And, like noses or eyes, these brain structures are not exactly identical from person to person.

Neuroscientists have long worked to describe and define particular regions of the brain. In some ways this is like describing parcels of land in property documents, and, like property descriptions, several different methods are used. At the largest scale, the brain is often divided into three parts: the brain stem, the cerebellum, and the cerebrum (Figure 3).⁷

7. The brain also is sometimes divided into the forebrain, midbrain, and hindbrain. This classification is useful for some purposes, particularly in describing the history and development of the

Figure 3. Lateral (left) and mid-sagittal (right) views of the human brain.



Source: Courtesy of Anthony Wagner, Stanford University.

The brain stem is found near the bottom of the brain and is, in some ways, effectively an extension of the spinal cord. Its various parts play crucial roles in controlling the body's autonomic (involuntary or unconscious) functions, such as heart rate and digestion. The brain stem also contains important regions that regulate processing in the cerebrum, one example of the strong pattern of interconnections in brain regions. For instance, the substantia nigra and ventral tegmental area in the brain stem consist of critical neurons that generate the neurotransmitter dopamine. While the substantia nigra is crucial for motor control, the ventral tegmental area is important for learning about rewards. The loss of neurons in the substantia nigra is associated with the movement problems of Parkinson's disease.

The cerebellum, which is about the size and shape of a squashed tennis ball, is tucked away in the back of the skull. It plays a major role in fine motor control and seems to keep a library of learned motor skills, such as riding a bicycle. Although it is relatively small by volume, it is the region of the brain with the largest number of neurons. It was long thought that damage to the cerebellum had little to no effect on a person's personality or cognitive abilities, but resulted primarily in unsteady gait, difficulty in making precise movements, and problems in learning movements. However, more recent studies of patients with cerebellar damage and functional brain imaging studies of healthy individuals indicate that

vertebrate brain, but it does not entirely correspond with the categorization of cerebrum, brain stem, and cerebellum, and it is not used in this reference guide.

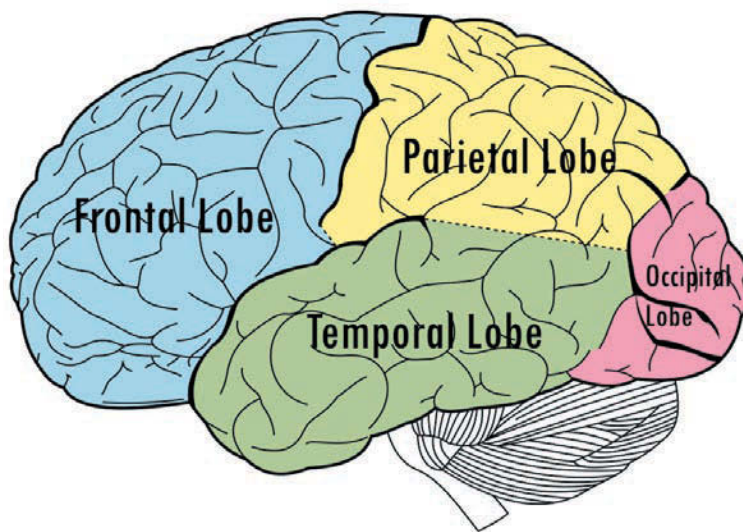
the cerebellum also plays a role in more cognitive functions, including supporting aspects of working memory, attention, and language.

The cerebrum is the largest part of the human brain, making up about 85% of its volume. The cerebrum is found at the front, top, and much of the back of the human brain. The human brain differs from the brains of other mammals mainly in that it has a vastly enlarged cerebrum.

There are several different ways to identify parts of, or locations in, the cerebrum. First, the cerebrum is divided into two hemispheres—the famous left and right brain. These two hemispheres are connected by tracts of white matter—of axons—most notably the large connection called the corpus callosum. Oddly, in humans and most other vertebrates, the right hemisphere of the brain generally receives messages from and controls the movements of the left side of the body, while the left hemisphere receives messages from and controls the movements of the right side of the body.

Each hemisphere of the cerebrum is divided into four lobes (Figure 4): the frontal lobe in the front of the cerebrum (behind the forehead), the parietal lobe at the top and toward the back, the temporal lobe on the side (just behind and above the ears), and the occipital lobe at the back. Thus, one could describe a particular region as lying in the left frontal lobe—the frontal lobe of the left hemisphere.

Figure 4. Lobes of a hemisphere. Each hemisphere of the brain consists of four lobes: the frontal, parietal, temporal, and occipital lobes.



Source: Adapted from https://commons.wikimedia.org/wiki/File:Lobes_of_the_brain_NL.svg.

The surface of the cerebrum consists of the cortex, which is a sheet of gray matter a few millimeters thick. The cortex is not a smooth sheet in humans, but rather is heavily folded with valleys, called sulci (“sulcus” in the singular), and bulges, called gyri (“gyrus”). The sulci and gyri have their own names, so a location can be described as in the inferior frontal gyrus in the left frontal lobe. These folds allow the surface area of the cortex, as well as the total volume of the cortex, to be much greater than in other mammals, while still allowing it to fit inside our skulls, similar to the way the many folds of a car’s radiator give it a very large surface area (for radiating away heat) in a relatively small space.

The cerebral cortex is extraordinarily large in humans compared with other species and is clearly centrally involved in much of what makes our brains special, but the cerebrum contains many other important subcortical structures that we share with other vertebrates. Some of the more important of these structures include the thalamus, the hypothalamus, the basal ganglia, and the amygdala. These areas all connect widely, with the cortex, with each other, and with other parts of the brain, to form complex networks.

The functions of all these areas are many, complex, and not fully understood, but some facts are known. The thalamus seems to act as a main relay that carries information to and from the cerebral cortex, particularly for vision, hearing, touch, and proprioception (one’s sense of the position of the parts of one’s body). It also is, importantly, involved in sleep, wakefulness, and consciousness. The hypothalamus has a wide range of functions, including the regulation of body temperature, hunger, thirst, and fatigue. The basal ganglia are a group of regions in the brain that are involved in motor control and learning, among other things. They seem to be strongly involved in selecting movements, as well as in learning through reinforcement (as a result of rewards). The amygdala appears to be important in emotional processing, including how we attach emotional significance to particular stimuli.

In addition, many other parts of the brain, in the cortex or elsewhere, have their own special names, usually with Latin or Greek roots that may or may not seem descriptive today. The hippocampus, for example, is named for the Greek word for seahorse. For most of us, these names will have no obvious rhyme or reason, but merely must be learned as particular structures in the brain—the superior colliculus, the tegmentum, the globus pallidus, the substantia nigra, the cingulate cortex, and more. All of these structures come in pairs, with one in the left hemisphere and one in the right hemisphere; only the pineal gland is unpaired. Brain atlases include scores of names for particular structures or regions in the brain and detailed information about those structures or regions.

Some of these smaller structures may have special importance to human behavior. The nucleus accumbens, for example, is a small subcortical region in each hemisphere of the cerebrum that appears important for reward processing and motivation. In experiments with rats that received stimulation of this region

in return for pressing a lever, the rats would press the lever almost to the exclusion of any other behavior, including eating. The nucleus accumbens in humans appears linked to appetitive motivation, responding in anticipation of primary rewards (such as pleasure from food or sex) and secondary rewards (such as money). Through interactions with the orbital frontal cortex and dopamine-generating neurons in the midbrain (including the ventral tegmental area), the nucleus accumbens is considered part of a “reward network.” With a hypothesized role in addictive behavior and in reward computations, more broadly, this putative reward network is a topic of considerable ongoing research.

All of these various locations, whether defined broadly by area or by the names of specific structures, can be further subdivided using directions: front and back, up and down, toward the middle or toward the sides. Unfortunately, the directions often are not expressed in a straightforward manner, and several different terminological conventions exist. Locations toward the front or back of the brain can be referred to as either anterior or posterior, or as rostral or caudal (literally, toward the nose, or beak, or toward the tail). Locations toward the bottom or top of the brain are termed inferior or superior or, alternatively, as ventral or dorsal (toward the stomach or toward the back). A location toward the middle of the brain is called medial; one toward the side is called lateral. Thus, different locations could be described, for example, as in the left anterior cingulate cortex, in the dorsal medial (or sometimes dorsomedial) prefrontal cortex, or in the posterior hypothalamus.

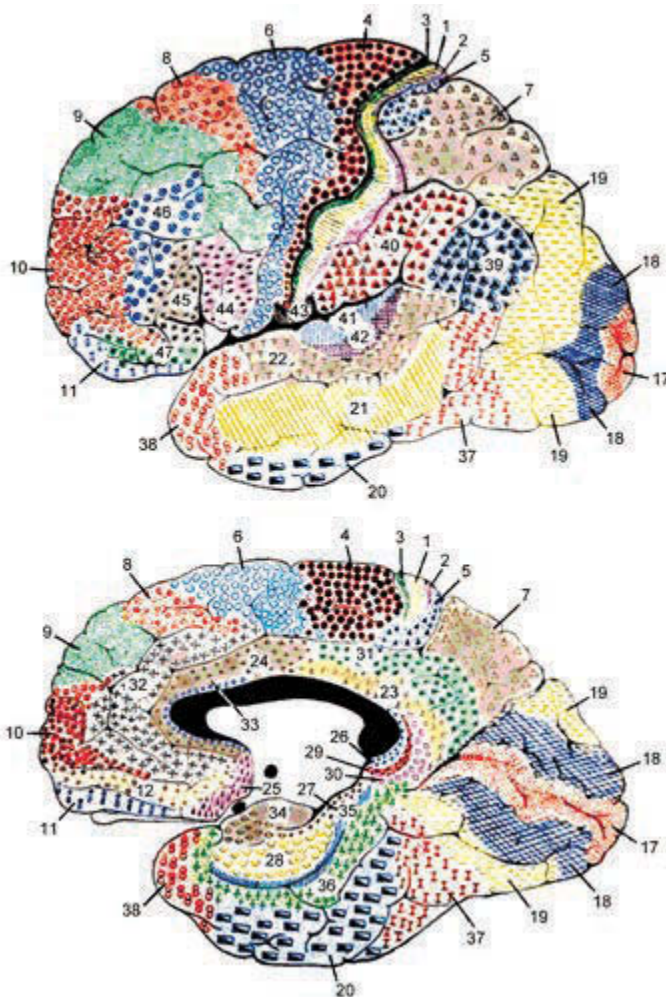
Finally, one other method often is used, a method created by Korbinian Brodmann in 1909. Brodmann, a neuroanatomist, divided the brain into about forty-five different areas or regions. Each region was defined based on the kinds of neurons found there and how those neurons are organized. A location described by a Brodmann area (Figure 5) may or may not correspond closely with a structural location. Other organizational schemes exist, but Brodmann’s remains the most widely used to describe the approximate locations of findings in modern human brain imaging studies.

If these varied ways of naming different brain regions seem confusing, it is because they are. But think of them as being like U.S. Supreme Court decisions being cited to *U.S. Reports*, the *Supreme Court Reporter*, and *Supreme Court Reports: Lawyers’ Edition*. They are different methods of pointing to particular parts of the brain.

Some Aspects of How the Brain Works

Most of neuroscience is dedicated to finding out how the brain works, but although much has been learned, considerably more remains unknown. We could use many different ways to describe what is known about how the brain

Figure 5. Brodmann's areas.



Source: Professor Mark Dubin, University of Colorado.

works. This section will discuss a few important aspects of brain function and will make several general points about the localization and distribution of functions, as well as brain plasticity, before commenting on the effects of hormones and other chemical influences on the brain.

Some brain functions are localized in, or especially dependent on, particular regions of the brain. This has been known for many years as a result of studies of people who, through traumatic injury, stroke, or cancer, have lost, or lost the use

of, particular regions of their brains. For example, in the 1860s, French anatomist Paul Broca discovered through autopsies of patients that damage to a region in the left inferior frontal lobe (now known as Broca's area) caused an inability to speak. It is now known that some functions cannot normally be performed when particular brain areas are damaged or missing. The visual cortex, located at the back of the brain in the occipital lobes, is as necessary for vision as the eyes are; the hippocampus is necessary for the creation of many kinds of memory; and the motor cortex is necessary for voluntary movements. The motor cortex and the parallel somatosensory cortex, which is essential for processing sensory information such as the sense of touch from the body, are further subdivided, with particular regions necessary for causing motion or sensing feelings from the legs, arms, fingers, face, and so on.

At the same time, the fact that a region is necessary to a particular class of sensations, behaviors, or cognition does not mean either that it is not involved in other brain functions or that other brain regions do not also contribute to these particular abilities. The amygdala, for example, is involved in our feelings of fear, but it is also involved broadly in emotional reactions, both positive and negative. It also modulates learning, memory, and even sensory perception. Although some functions are localized, others are widely distributed. For example, the visual cortex is essential to vision, but actual visual perception involves many parts of the brain in addition to the occipital lobes. Memories appear to be stored over much of the cortex. Networks of brain regions participate in many of these functions.

For example, if you touch something very hot with your left index finger, your spinal cord, through a reflex loop, will cause you to pull your finger back very quickly. Then the part of your right somatosensory cortex devoted to the index finger will be involved in receiving and initially interpreting the sensation. Other areas of your brain will recognize the stimulus as painful, your motor regions will be involved in waving your hand back and forth or bringing your finger to your mouth, widespread parts of your cortex may lead to you remembering other instances of burning yourself, and your hippocampus may play a role in making a new long-term memory of this incident. There is no single brain region dedicated to reacting when your finger burns; many regions, both specific and general, contribute to the brain's response.

In addition, brains are at least somewhat "plastic" or changeable on both small and large scales. Anyone who can see has a working visual cortex, and it is always located in the back of the brain (in the occipital lobe), but its exact borders will vary slightly from person to person. Even within individuals, the shape of the brain and the size and location of its parts can change over time. This is most evident before birth, during childhood, and for many people well into their twenties. It is not true that adults make no new neurons, but we do not make many of them, and they are localized to two regions: the olfactory bulb, which plays a huge role in our senses of smell and taste, and the hippocampus, which is

essential for making most kinds of memories. Unfortunately, though, adults can and do lose neurons through the shrinkage or even destruction of parts of the brain, as a result of traumatic injury, disease, or aging.

Another kind of plasticity is ubiquitous in our brains' lives. The shape or number of our neurons does not change much after childhood, but their connections and functions can change markedly. The brain can adjust and change in response to a person's behavior or changes in that person's anatomy. If a person loses an arm to amputation, the parts of the motor and somatosensory cortices that had dealt with that arm may be "taken over" by other body parts. In some cases, this brain plasticity can be extreme. Young children who have lost an entire hemisphere of their brains may grow up to have normal or nearly normal functionality as the remaining hemisphere takes on the tasks of the missing hemisphere. Unfortunately, the possibilities of this kind of extreme plasticity do diminish with age, but rehabilitation after stroke in adults sometimes shows changes in the brain functions undertaken by particular brain regions.

Both kinds of plasticity—physical changes in the brain and connective and functional changes in the brain—pose some important problems for neuroimaging evidence. A criminal defendant's brain might have unusual and possibly detrimental physical or functional anomalies at or shortly before the time of trial, but those issues may not have existed at the time of the relevant events. This "time machine" problem is discussed further at pages 1244 and 1249.

The picture of the brain as a set of interconnected neurons that fire in networks or patterns in response to stimuli is useful but not complete. In addition to neuron firings, other factors affect how the brain works, particularly chemical factors.

Some of these factors are hormones, generated by the body either inside or outside the brain. They can affect how the brain functions, as well as how it develops. Sex hormones such as estrogen and testosterone can have both short-term and long-term effects on the brain. So can other hormones, such as cortisol, associated with stress, and oxytocin, associated with, among other things, trust and bonding. Endorphins, chemicals secreted by the pituitary gland in the brain, are associated with pain relief and a sense of well-being. Still other chemicals, brought in from outside the body, can have major effects on the brain, both in the short term and the long term. Examples include alcohol, caffeine, nicotine, morphine, and cocaine. These can trigger very specific brain reactions or can have broad effects.

Some Common Neuroscience Techniques

Neuroscientists use many techniques to study the brain. Some of them have been used for centuries, such as autopsies and the observation of patients with

brain damage. Some, such as the intentional destruction of parts of the brain, can be used ethically only in research on nonhuman animals. Of course, research with nonhuman animals, while often helpful in understanding human brains, is of less value when examining behaviors that are uniquely developed among humans. The current revolution in neuroscience is largely the result of a revolution in the tools available to neuroscientists, as new methods have been developed to image and to intervene in living brains. These methods, particularly the imaging methods that allow more precise measurements of human brain structure and function in living people, are giving rise to increasing efforts to introduce neuroscientific evidence in court.

This section will focus on several kinds of neuroimaging—computerized axial tomography (CAT scans), positron emission tomography (PET scans), single photon emission computed tomography (SPECT scans), magnetic resonance imaging (MRI), diffusion tensor imaging (DTI), and functional near-infrared spectroscopy (fNIRS), as well as an older method, electroencephalography (EEG) and its close relative magnetoencephalography (MEG). Some of these methods show the structure of the brain, others show the brain's functioning, and some do both. These are not the only important neuroscience techniques; several others will be discussed briefly at the end of this section. Genetic analysis provides yet another technique for increasing our understanding of human brains and behaviors, but this reference guide will not deal with the possible applications of human genetics to understanding behavior.

We go into these techniques at some length and in some detail, but we think that is crucial. A judge may be confronted with neuroscience evidence that has been developed using any of these methods. Having at least some background in what they are and their strengths and weaknesses may be invaluable. In the context of any given case, a judge may well choose to look at only the parts of this section relevant to that case rather than read the whole section.

Before digging into the techniques, though, it may be useful to discuss briefly the experts who will be testifying about them. All scientific evidence is presented through experts. In court cases involving brain or mind evidence, judges and juries regularly hear testimony from clinical psychologists and psychiatrists, often specialized as forensic psychologists or psychiatrists. Other kinds of specialist or subspecialist physicians may also testify, such as neurologists, neurosurgeons, neurointensivists, neuroradiologists, or neuropathologists.

Those experts may also be involved in presenting neuroscience evidence, but they will likely be joined or replaced by nonclinical experts in neuroscience, usually from universities or medical schools. Some will be located in departments of neuroscience, neurobiology, or some other term their employer uses for its basic brain science research department; others will be in psychology, psychiatry, neurology, or neurosurgery departments. Although some of these experts will be trained physicians and may even see some patients, most will have doctoral degrees and will focus their work on research. Rather than testifying about the condition

or history of a particular person, they will usually testify about neuroscience techniques or research establishing connections between the findings of those techniques and relevant human behaviors or mental conditions, based on peer-reviewed literature, often that they have published. They may or may not have examined the individual in question.

The distinctions among the types of experts who might testify about neuroscience findings is well discussed in a 2018 law review article.⁸

Neuroimaging

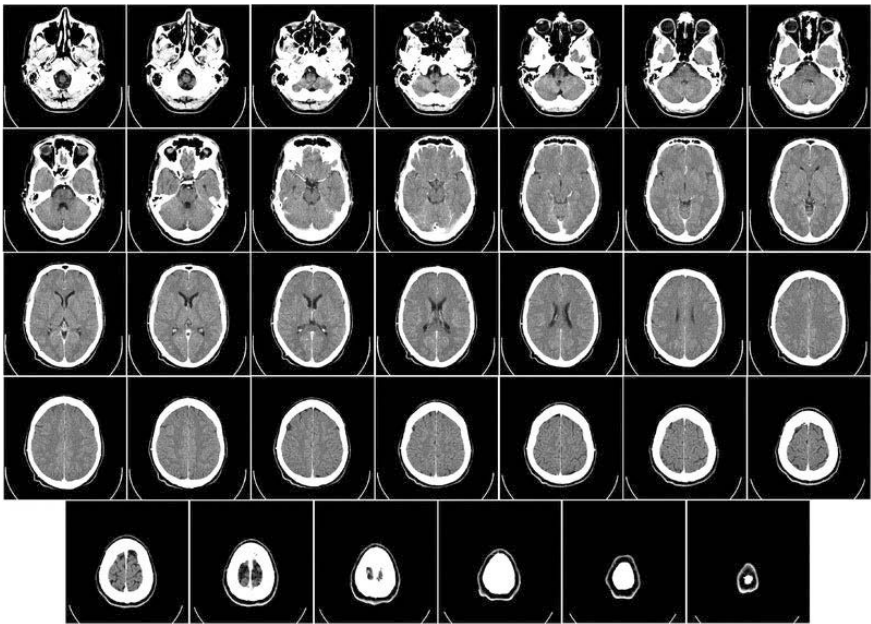
Although X-ray imaging has been used for over a century, it has not been very helpful in studying the brain. X-ray images are the shadows cast by dense objects. Not only is the brain surrounded by our very dense skulls, but there are no dense objects inside the brain to cast these shadows. Although a few features of the brain or its blood vessels could be seen through methods that involved the injection of air into some of the spaces in the brain or of contrast media into the blood, these provided limited information. The opportunity to see inside a living brain itself only goes back to about the 1970s, with the development of CAT scans. This ability has since exploded with the development of several new techniques. This section will discuss five of them: CAT scans (sometimes also called computer-assisted tomography or computed tomography), PET and SPECT scans, MRI, DTI, and fNIRS.

CAT Scans

The CAT scan is a multidimensional, computer-assisted X-ray machine. Instead of taking one X-ray from a fixed location, in a CAT scan both the X-ray source and the X-ray detectors opposite the source rotate around the person being scanned. Traditional X-ray machines shine the radiation through an object to a photographic negative, producing an actual picture that shows where dense objects blocked the X-rays. In a CAT scan (Figure 6), the X-ray detectors produce data sufficient to reconstruct the scanned object in three dimensions. Computerized algorithms can then be used to produce an image of any particular slice through the object. The multiple angles and computer analysis make it possible to pick out the relatively small density differences within the brain that traditional X-ray technology could not distinguish and to use them to produce images of the soft tissue.

8. See Jane Campbell Moriarty & Daniel D. Langleben, *Who Speaks for Neuroscience? Neuroimaging Evidence and Courtroom Expertise*, 68 Case W. Rsrv. L. Rev. 783 (2018), <https://perma.cc/ECK8-D7WS>.

Figure 6. CAT scan depicting axial sections of the human brain.



Source: https://en.wikipedia.org/wiki/File:Computed_tomography_of_human_brain_-_large.png.

The CAT scan provides a structural image of the brain. It is useful for showing some kinds of structural abnormalities, but it provides no direct information about the brain's functioning. A CAT scan brain image is not as precise as the image produced from an MRI, but because the procedure is both quick and (relatively) inexpensive, CAT scanners are common in hospitals. Medically, brain CAT scans are used mainly to look for bleeding or swelling inside the brain, although they also will record sizable tumors or other large structural abnormalities. For neuroscience, the great advantage of the CAT scan was its ability, for the first time, to see some details inside the skull, an ability that has been largely superseded for research by MRI. CAT scans have been used for decades in litigation to provide evidence about the physical condition of the brain, often in cases about brain injuries or medical malpractice. But they have also been used to argue that structural changes in the brain, shown on the CAT scan, are evidence of insanity or other mental impairments. Perhaps their most notable use was in 1982 in the trial of John Hinckley Jr., for the attempted assassination of President Ronald Reagan. A CAT scan of Hinckley's brain that showed widened sulci (the "valleys" in the surface of the brain) was introduced

into evidence to show that Hinckley suffered from organic brain damage in the form of shrinkage of his brain.⁹

PET Scans and SPECT Scans

Traditional X-ray machines and their more sophisticated descendant, the CAT scanner, project X-rays through the skull and create images based on how much of the X-rays are blocked or absorbed. PET scans and SPECT scans operate very differently. In these methods, a substance that emits radiation is introduced into the body. That radiation then is detected from outside the body in a way that can determine the location of the radiation source. These scans generally are not used for determining the brain's structure, but are used for understanding how it is functioning. They are particularly good at measuring one aspect of brain structure—the density of particular receptors, such as those for dopamine, at synapses in some areas of the brain, such as the frontal lobes.

Radioactive decay of atoms can take several forms, producing alpha, beta, or gamma radiation. PET scanners take advantage of isotopes of atoms that decay by giving off positive beta radiation. Beta decay usually involves the emission of an electron; positive beta decay involves the emission of a positron, the positively charged antimatter equivalent of an electron. When positrons (antimatter) meet electrons (matter), the two particles are annihilated and converted into two photons of gamma radiation with a known energy (511,000 electron volts) that follow directly opposite paths from the site of the annihilation. Inside the body, the collision between the positron and electron and the consequent production of the gamma radiation photons take place within a short distance (a millimeter or two) of the site of the initial radioactive decay that produced the positron.

PET scans, therefore, start with the introduction into a person's body of a radioactive tracer that decays by giving off a positron. One common tracer is fluorodeoxyglucose (FDG), a molecule that is almost identical to the simple sugar glucose, except that one of the oxygen atoms in glucose is replaced by an atom of fluorine-18, an isotope of the element fluorine with nine protons and nine neutrons. Fluorine normally found in nature is fluorine-19, with nine protons and ten neutrons, and is stable. Fluorine-18 is very unstable and decays quickly, through positive beta decay, losing about half of its mass every 110 minutes (its half-life). The body treats FDG as though it were glucose, and so the FDG is concentrated where the body needs the energy supplied by glucose. A major clinical

9. The effects of this evidence on the verdict are unclear. See Lincoln Caplan, *The Insanity Defense and the Trial of John W. Hinckley, Jr.* (1984), for a discussion of the case and its consequences for the law.

use of PET scans derives from the fact that tumor cells use energy, and hence glucose, at much higher rates than normal cells.

After giving the FDG time to become concentrated in the body, which usually takes about an hour, the person is put inside the scanner itself. There, the person is entirely surrounded by a very sensitive radiation detector, tuned to respond to gamma radiation of the energy produced by annihilated positrons. When two “hits” are detected by two sensors at about the same time, the source is known to be located on a line connecting the two. Very small differences in the timing of when the radiation is detected can help determine where along that line the annihilation took place. In this way, as more gamma radiation from the decaying FDG is detected, the general location of the FDG within the body can be determined and, as a result, tissue that is using a lot of glucose, such as a tumor, can be located.

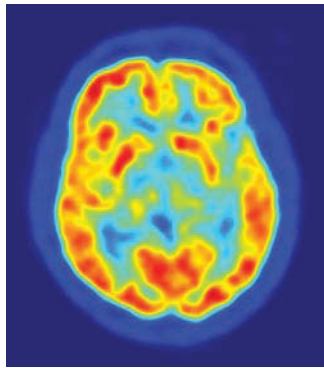
In neuroscience research, PET scans also can be taken using different molecules that bind more specifically to particular tissues or cells. Some of these more specific ligands use fluorine-18, but others use a different radioactive tracer that also decays by emitting a positron—oxygen-15. This can be used to determine what parts of the brain are using more or less oxygen. Oxygen-15, however, has a much shorter half-life (two minutes) and so is more difficult and expensive to use than FDG. Similarly, carbon-11, with a half-life of twenty minutes, also can be used. Carbon-11 atoms can be introduced into various molecules that bind to important receptors in the brain, such as receptors for dopamine, serotonin, or opioids. This allows the study of the distribution and function of these receptors, both in healthy people and in people with various mental illnesses or neurological diseases.

The result of a PET scan is a record of the locations of positron decay events in the brain. Computer visualization tools can then create cross-sectional images of the brain, showing higher and lower rates of decay, with differences in magnitude typically depicted through the use of different colors.

PET scans are excellent for showing the location of various receptors in normal and abnormal brains (Figure 7). PET scans are also very good for showing areas of different glucose use and, hence, of different levels of metabolism. This can be very useful, for example, in detecting some kinds of brain damage, such as the damage that occurs with Alzheimer’s disease, where certain regions of the brain become abnormally inactive, or in brain regions that have been damaged by a stroke.

In addition, the comparison (subtraction) of two PET scan measurements—one scan when a person is engaged in a task that is thought to require particular brain functions and a second control (or baseline) scan that is not thought to require these functions—allows researchers indirectly to measure brain function. PET scans were initially used in this way in research to show what areas of the brain were used when people experienced various stimuli or performed particular tasks. It has been substantially superseded for this purpose by functional

Figure 7. PET scan depicting an axial section of the human brain.



Source: <https://commons.wikimedia.org/wiki/File:PET-image.jpg>.

MRI, which is less expensive, does not involve radiation exposure, provides better spatial resolution, and allows a longer period of testing.

SPECT scans are similar to PET scans. Each can produce a three-dimensional model of the brain and display images of any cross section through the brain. Like PET scans, they require the injection of a radioactive tracer material; unlike PET scans, the radioactive tracer directly emits gamma radiation rather than emitting positrons. These kinds of tracers are more stable, more accessible, and much cheaper than the positron-emitting tracers needed for PET scans. With a PET scan, the gamma detector entirely surrounds the person; with a SPECT scan, one to three gamma detectors are rotated around the body for about fifteen to twenty minutes. As with PET scans, the SPECT tracers can be used to measure brain metabolism or to attach to specific molecular receptors in the brain. The spatial resolution of a SPECT scan, however, is poorer than with a PET scan, with an uncertainty of about one centimeter.

Both PET and SPECT scans are most useful if coupled with good structural images. Contemporary PET and SPECT scanners often include a simultaneous CAT scan; there is some experimental work aimed at providing simultaneous PET and MRI scans.

MRI—Structural and Functional

MRI was developed in the 1970s, first came into wide use in the 1980s, and is currently the dominant neuroimaging technology for producing detailed images of the brain's structure and for measuring aspects of brain function. MRI

operates on completely different principles than either CAT scans or PET or SPECT scans; it does not rely on X-rays passing through the brain or on the decay of radioactive tracer molecules inside the brain. Rather, MRI's workings involve more complicated physics. This section will discuss the general characteristics of magnetic resonance imaging and then will focus on structural MRI, diffusion tensor imaging, and finally functional MRI.

The power of an MRI scanner is measured by the strength of its magnetic field, measured in units called tesla (T). The magnetic field of a small bar magnet is about 0.01T. The strength of the Earth's magnetic field is about 0.00005T. The MRI machines used for clinical purposes use magnetic fields of between 0.2T and 3.0T, with 1.5T or 3.0T being the systems most commonly used today. MRI machines for human research purposes have reached 9.4T. In general, the stronger the magnetic field, the better the image, although higher fields also can create their own measurement difficulties, especially when imaging brain function. MRI machines achieve these high magnetic fields through use of superconducting magnets, made by cooling the electromagnet with liquid helium at a temperature four degrees Celsius above absolute zero. For this and other reasons, MRI systems are complicated, with higher initial and continuing maintenance costs compared to some other methods for functional imaging (e.g., electroencephalography; see below).

In most MRI systems, the individual, on an examination table, slides into a cylindrical opening in the machine so that the part of the body to be imaged is in the middle of the magnet (Figure 8). Depending on the kind of imaging performed, the examination or experiment can take from about thirty minutes to more than two hours; throughout the scanning process the individual being examined needs to stay as motionless as possible to avoid corrupting the images. The main sensations are the loud thumping and buzzing noises made by the machine, as well as the machine's vibration.

MRI examinations appear to involve minimal risk. Unlike the other neuroimaging technologies discussed above, MRI does not involve any high-energy radiation. The magnetic field seems to be harmless, at least as long as magnetizable objects are kept away from it. MRI participants need to remove most metal objects; people with some kinds of implanted metallic devices, with tattoos with metal in their ink, or with fragments of ferrous metal anywhere in their bodies cannot be scanned because of the dangerous effects of the field on those bits of metal.

When an individual is positioned in the MRI scanner, the powerful field of the magnet causes the nuclei of atoms (usually the hydrogen nuclei of the body's water molecules) to align with the direction of the main magnetic field of the magnet. Using a brief electromagnetic pulse, these aligned atoms are then "flipped" out of alignment from the main magnetic field, and, after the pulse stops, the nuclei then rapidly realign with the strong main magnetic field. Because the

Figure 8. MRI machine. Magnetic resonance imaging systems are used to acquire both structural and functional images of the brain.



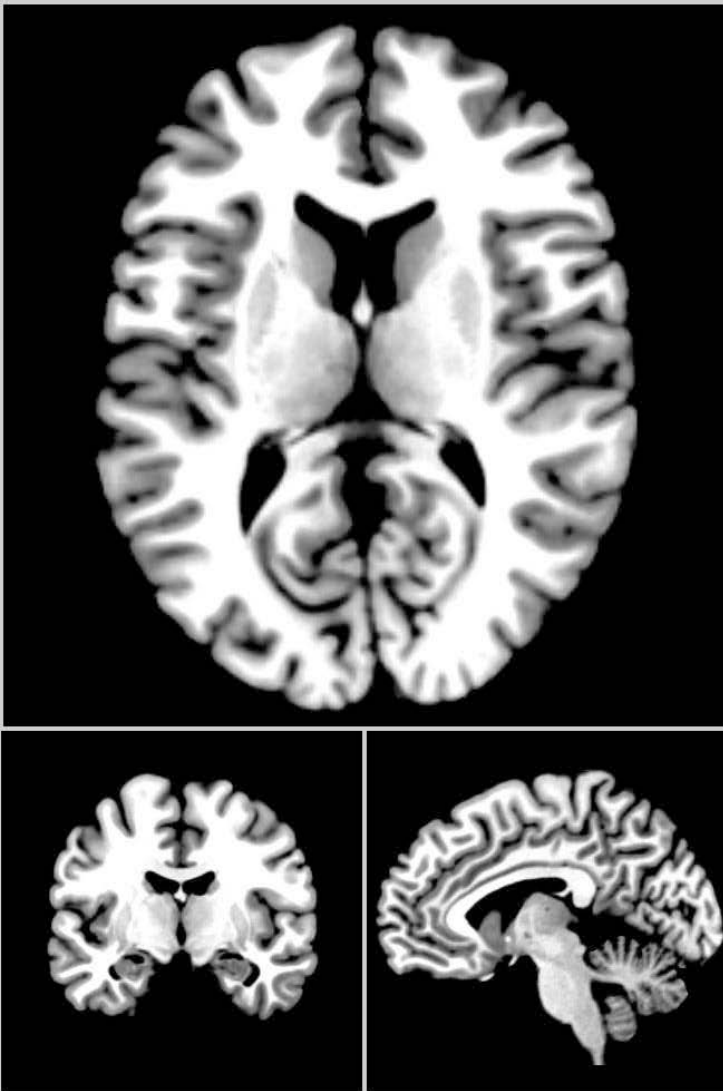
Source: Courtesy of Anthony Wagner, Stanford University.

nuclei spin (like a top), they create an oscillating magnetic field that is measured by a receiver coil. During structural imaging, the strength of the signal generated partially depends on the relative density of hydrogen nuclei, which varies from point to point in the body according to the density of water. In this manner, MRI scanners can generate images of the body's anatomy or of other scanned objects. Because an MRI scan can effectively distinguish between similar soft tissues, MRI can provide very high-resolution images of the brain's anatomy, which is, after all, made up of soft tissue.

Structural MRI scans produce very detailed images of the brain (Figure 9). They can be used to spot abnormalities, large and small, as well as to see normal

Reference Guide on Neuroscience

Figure 9. Brain MRI scan depicting an axial (upper), coronal (lower left), and sagittal (lower right) image of the human brain.



Source: Courtesy of Anthony Wagner, Stanford University.

variation in the size and shape of brain features. Structural MRI can be used, for example, to see how brain features change as a person ages. Previously, getting that kind of detailed information about a brain required an autopsy or, at a minimum, extensive neurosurgery. This ability makes structural MRI both an important clinical tool and a very useful technique for research that tries to correlate human differences, normal and abnormal, to differences in brain structure, as well as for research that seeks to understand brain development.

Functional MRI (fMRI; Figure 10) harnesses the power of MRI in neuroscience for understanding brain function. While it is perhaps one of the most exciting technologies used in neuroscience, as discussed in the section titled “Replication” below, it has proven difficult to replicate many findings from fMRI research. The technique shows what regions of the brain are more or less active in response to the performance of particular tasks or the presentation of particular stimuli. It does not measure brain activity (the firing of neurons) directly but, instead, looks at how blood flow changes in response to brain activity and uses those changes, through the so-called BOLD response (the blood-oxygen-level-dependent response), to allow the researcher to infer patterns of brain activity.

Structural MRI generally creates its images through detecting the density of hydrogen atoms in the participant and flipping them with radio pulses. For fMRI, the scanner detects changes in the ratio of oxygenated hemoglobin (oxyhemoglobin) and deoxygenated hemoglobin (deoxyhemoglobin) in particular locations of the brain. Hemoglobin is the protein in red blood cells that carries oxygen from the lungs to the body. Based on metabolic demands, hemoglobin molecules supply oxygen for the body’s needs. Accordingly, “fresher” blood will have a higher ratio of oxyhemoglobin to deoxyhemoglobin than more “used” blood. Importantly, because deoxyhemoglobin (which is found at a higher level in “used” blood) causes the fMRI signal to decay, a higher ratio of oxyhemoglobin to deoxyhemoglobin will produce a stronger fMRI signal.

Neural activity is energy intensive for neurons, and neurons do not contain any significant reserves of oxygen or glucose. Therefore, the brain’s blood vessels respond quickly to increases in activity in any one region of the brain by sending more fresh blood to that area. This is the basis of the BOLD response, which measures changes in the ratio of oxyhemoglobin to deoxyhemoglobin in a brain region several seconds after activity in that region. In particular, when a brain region becomes more active, there is first, perhaps more intuitively, a decline in the ratio of oxyhemoglobin to deoxyhemoglobin immediately after activity in the region, apparently corresponding to the depletion of oxygen in the blood at the site of the activity. This decline, however, is very small and very hard to detect with fMRI. Immediately after this decrease, there is an infusion of fresh (oxyhemoglobin-rich) blood, which can take several seconds to reach maximum; it is this infusion that results in the increase in the oxy/deoxyhemoglobin ratio

that is measured in BOLD fMRI studies. Because even this subsequent increase is relatively small and variable, fMRI experiments typically involve many trials of the same task or class of stimuli in order to be able to see the signal amid the noise.

Thus, in a typical fMRI experiment, the participant will be placed in the scanner and the researchers will measure differences in the BOLD response throughout their brain under different conditions. A participant might, for example, be told to look at a video screen on which images of places alternate with images of faces. For purposes of the experiment, the computer will impose a spatial map on the participant's brain, dividing it into thousands of little cubes, each a few cubic millimeters in size, referred to as "voxels." Either while the data are being collected (so-called real-time fMRI¹⁰) or after an entire dataset has been gathered, a computerized program will compare the BOLD signal for each voxel when the screen was showing places to the signals when the screen contained faces. Regions that showed a statistically significant increase in the BOLD response several seconds after the face was on the video screen compared with the effects several seconds after a screen showing a face appeared. The researchers will infer that those regions were, in some way, involved in how the brain processes images of faces. The results will typically be shown as a structural brain image on which areas of more or less activation, as shown by a statistical test, will be represented by different colors.¹¹

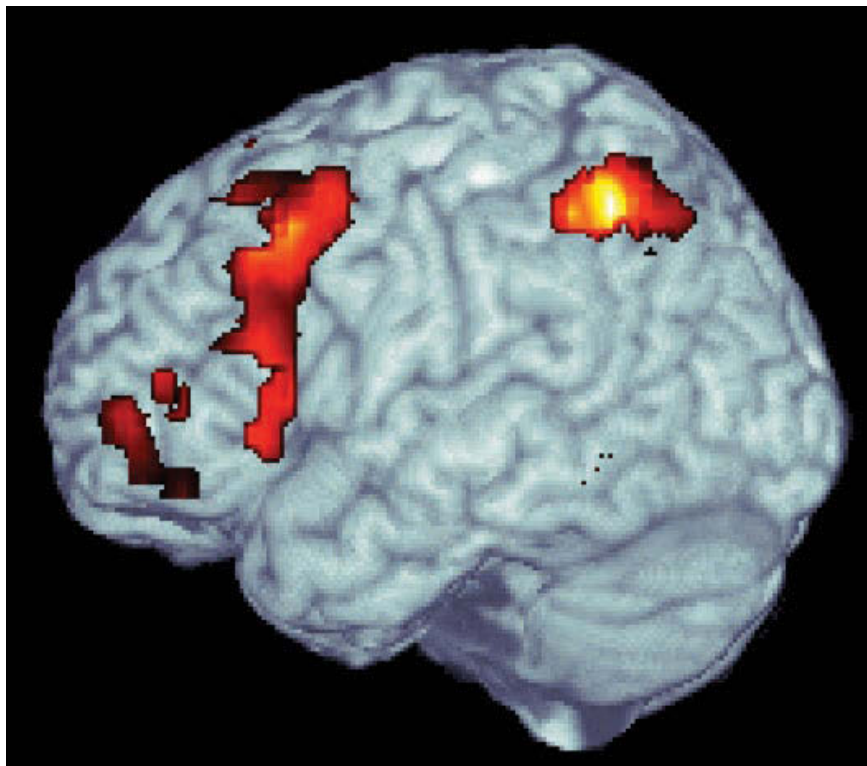
fMRI was first proposed in 1990, and the first research results using BOLD-contrast fMRI in humans were published in 1992. The first decade of this century saw an explosive increase in the number of research articles based on fMRI, with nearly 2,500 articles published in 2008 alone—compared to about 450 in 1998.¹² The tide continued to rise; in 2015 alone, nearly 30,000 fMRI papers were published. MRI (functional and structural) is quite safe, and MRI machines are widespread in developed countries, largely for clinical use but increasingly for research use as well. Although fMRI research is subject to many questions and controversies (discussed below), this technique sparked a surge of

10. Use of this "real-time fMRI" has been increasing, but it is not yet clear whether the claims for it will stand up.

11. This example is actually a simplified version of experiments performed by Professor Nancy Kanwisher at MIT in the early 2000s that explored a region of the brain called the fusiform face area that is particularly involved in processing visions of faces. See Kathleen M. O'Craven & Nancy Kanwisher, *Mental Imagery of Faces and Places Activates Corresponding Stimulus-Specific Brain Regions*, 12 *J. Cognitive Neuroscience* 1013 (2000), <https://doi.org/10.1162/08989290051137549>.

12. For a census of fMRI articles from 1993 to 2008, see Carole A. Federico, Sofia Lombera & Judy Illes, *Intersecting Complexities in Neuroimaging and Neuroethics*, in *Oxford Handbook of Neuroethics* 377 (Judy Illes & Barbara J. Sahakian eds., 2012), <https://doi.org/10.1093/oxfordhb/9780199570706.013.0098>. This continued an earlier census from 1993 to 2001. Judy Illes, Matthew P. Kirschen & John D.E. Gabrieli, *From Neuroimaging to Neuroethics*, 6 *Nature Neuroscience* 205 (2003), <https://doi.org/10.1038/nn0303-205>.

Figure 10. fMRI image. Functional MRI data reveal regions associated with cognition and behavior. Here, regions of the frontal and parietal lobes that are more active when remembering past events relative to detecting novel stimuli are depicted.



Source: Courtesy of Anthony Wagner, Stanford University.

interest and commentary on the relationship between law and neuroscience, from questions about criminal responsibility and lie detection to how judges¹³ and jurors¹⁴ make decisions.

13. See, e.g., Anna Spain Bradley, *The Disruptive Neuroscience of Judicial Choice*, 9 U.C. Irvine L. Rev. 1 (2018), <https://perma.cc/UVL3-EJEA> (drawing from neuroscience evidence, the author argues that rather than based purely on rationality, judicial decision-making is influenced by bias, emotion, and empathy).

14. Edith Greene & Brian S. Cahill, *Effects of Neuroimaging Evidence on Mock Juror Decision Making*, 30 Behav. Scis. & L. 280 (2012), <https://doi.org/10.1002/bsl.1993>; Shelby Hunter, N. J. Schweitzer & Jillian M. Ware, *Neuroscience and Jury Decision-Making*, in *Criminal Juries in the*

DTI

DTI (diffusion tensor imaging) is a different application of MRI, one that uses the MRI data not to build a picture of the brain (structural MRI) or of changes in the ratio of oxygenated to deoxygenated hemoglobin over time in the brain (functional MRI) but to follow the ways water diffuses through brain tissue to map out the brain's white matter. As noted above, neuronal tissue in the brain can be divided roughly into gray matter (the bodies of neurons) and white matter (neuronal axons that transmit signals over distance). DTI uses MRI to see what direction water diffuses through brain tissue. Tracts of white matter are made up of bundles of axons coated with fatty myelin. Water will diffuse through that white matter along the direction of the axons and not, generally, across them. This method can be used, therefore, to trace the location of these bundles of white matter and hence the long-distance connections between different parts of the brain.

These kinds of “wiring diagrams” can be very important in brain research, but they also have clinical applications. Abnormal patterns of these connections may be associated with various conditions, from traumatic brain injury and stroke to brain tumors, Alzheimer's disease, and dyslexia, some of which may have legal implications.¹⁵

fNIRS

Like functional magnetic resonance imaging (fMRI), functional near-infrared spectroscopy can detect changes in hemoglobin within the brain, but it does so by measuring differences in the absorption of light that occurs through blood flow to different regions of the brain. fNIRS is effective because biological tissue is relatively transparent to infrared light in wavelengths ranging from 650 to 925 nanometers (nm). Light in this range is absorbed more strongly by oxygenated versus deoxygenated blood, and at different wavelengths. Deoxygenated blood is absorbed more strongly below 790 nm, and oxygenated blood more strongly above 790 nm.¹⁶ By detecting the changes in the relative concentrations of

21st Century: Psychological Science and the Law 221 (Cynthia Najdowski & Margaret Stevenson eds., 2018), <https://doi.org/10.1093/oso/9780190658113.003.0011>.

15. See Jennifer Christine van Velkinburgh, Mark D. Herbst & Stewart M. Casper, *Diffusion Tensor Imaging in the Courtroom: Distinction Between Scientific Specificity and Legally Admissible Evidence*, 11 World J. Clinical Cases 4477 (2023), <https://perma.cc/LLV4-28L3>; Andrew M. Lehmkuhl II, *Diffusion Tensor Imaging: Failing Daubert and Fed. R. Evid. 702 in Traumatic Brain Injury Litigation*, 87 U. Cin. L. Rev. 279 (2018), <https://perma.cc/GW32-7BUX>.

16. Wei-Liang Chen et al., *Functional Near-Infrared Spectroscopy and Its Clinical Application in the Field of Neuroscience: Advances and Future Directions*, 14 Frontiers Neuroscience 1 (2020), <https://doi.org/10.3389/fnins.2020.00724>.

different light-absorbing molecules, fNIRS can measure energy metabolism in the brain, including regional changes in brain activity.

fNIRS is portable, noninvasive, and cost-effective, and it allows real-time temporal tracking of brain activity. For these reasons, it is often used in clinical and research-based settings. For example, in the pediatric ICU, brain oxygenation levels are regularly monitored using fNIRS where long-term real-time monitoring of brain activity is necessary.¹⁷ Its disadvantages are that it cannot be used on parts of the brain more than four centimeters (a bit more than one and a half inches) below the skull and that its spatial resolution is low compared to fMRI, but higher than EEG, which we turn to next.

EEG and MEG

EEG (electroencephalography) is the measurement of the brain's electrical activity as exhibited on the scalp; MEG (magnetoencephalography) is the measurement of the small magnetic fields generated by the brain's electrical activity. The roots of EEG go back to the nineteenth century, but its use increased dramatically in the 1930s and 1940s.

The process for medical or research-grade EEG uses electrodes touching the participant's head, applied with an electrically conductive substance (a paste or a gel) to record electrical currents on the surface of the scalp. Multiple electrodes are used; for clinical purposes, 16 to 25 electrodes are commonly used, although arrays of more than 200 electrodes can be used. In MEG, superconducting "SQUIDS"¹⁸ are positioned over the scalp to detect the brain's tiny magnetic signals. The electrical currents are generated by the neurons throughout the brain, although EEG is more sensitive to currents emerging from neurons closer to the skull. It is therefore more challenging to use EEG to reveal the functioning of structures deep in the brain.

Because EEG and MEG directly measure neural activity, in contrast to the measures of blood flow in fMRI and fNIRS, the timing of the neural activity can be measured with great precision (the temporal resolution), down to milliseconds. On the other hand, in comparison to fMRI, EEG and MEG are poor at determining the location of the sources of the currents (the spatial resolution). Any one pattern of EEG or MEG signal at the scalp has an infinite number of possible source patterns, making the problem of determining the brain source of measured EEG/MEG signal particularly challenging and the results less precise.

17. *Id.*

18. SQUID stands for superconducting quantum interference device (and has nothing to do with the marine animal). This device can measure extremely small magnetic fields, including those generated by various processes in living organisms, and so is useful in biological studies.

Thus, EEG/MEG signal is a summation of the activity of thousands to millions of neurons at any one time.

The results of clinical EEG and MEG tests can be very useful for detecting some kinds of brain conditions, notably epilepsy, and are also part of the process of diagnosing brain death. EEG and MEG are also used for research, particularly in the form of event-related potentials, which correlate the size or pattern of the EEG or MEG signal with the performance of particular tasks or the presentation of particular stimuli. Thus, as with the hypothetical fMRI experiment described above, one could look for any consistent changes in the EEG or MEG signal when a participant sees faces rather than a blank screen. Apart from the determination of brain death, where EEG is already used, the most discussed possible legally relevant uses of EEG have been lie detection and memory detection.

EEG is safe, cheap, quiet, and portable. MEG is safe and quiet, but the technology is considerably more expensive than EEG and is not easily portable. EEG methods can tolerate much more head movement by the participant than PET or MRI techniques, although movement is often a challenge for MEG. EEG and MEG have good temporal resolution, distinguishing between milliseconds, which makes them very attractive for research, but their spatial resolution is inadequate for many research questions. As a result, some researchers use a combination of methods, integrating MRI and EEG or MEG data (acquired simultaneously or at different times) using sophisticated data analysis techniques.

Although EEG has most often been used in clinical and research-based settings, an increasing number of companies offer consumer-based devices with EEG electrodes. These devices have many fewer electrodes than clinical-grade EEG devices, and so offer much lower resolution and much noisier data. Large mainstream technology companies from Meta to Microsoft are developing consumer-based neural interface devices to enable consumers to meditate, play brain-controlled games, monitor their brain health, or even one day replace peripheral devices like keyboards and mice to interact with other technology. Widespread adoption of these consumer-based neural interface devices could significantly increase the amount of neural data being passively recorded in everyday life and potentially introduced into the legal system. The quality of the data and algorithms used to process the collected data will become increasingly at issue as these data are introduced into legal settings.

Other Techniques

Several other neuroscience techniques may also have legal applications. This section will briefly describe four other methods that may be discussed in court: lesion studies, transcranial magnetic stimulation, deep brain stimulation, and implanted microelectrode arrays.

Lesion Studies

One powerful way to test whether particular brain regions are associated with particular mental processes is to study mental processes after those brain regions have been destroyed or damaged. Observations of the consequences of such lesions, created by accidents or disease, were, in fact, the main way in which localization of brain function was originally understood.

For ethical reasons, the experimental destruction of brain tissue is limited to nonhuman animals. Nonetheless, in addition to accidental damage, on occasion, human brains will need to be intentionally damaged for clinical purposes. Tumors may have to be removed or, in some cases, epilepsy may have to be treated by removing the region of the brain where the seizures began. Valuable knowledge may be gained from following these individuals.

Our understanding of the role of the hippocampus in creating memories, as one example, was greatly aided by study of a patient known as H.M.¹⁹ When he was twenty-seven years old, H.M. was treated for intractable epilepsy, undergoing an experimental procedure that surgically removed his left and right medial temporal lobes, including most of his two hippocampi. The surgery was successful, but from that time until his death in 2008, H.M. could not form new long-term memories, either of events or of facts. His short-term memory, also known as working memory, was intact, and he could learn new motor, perceptual, and (some) cognitive skills (his “procedural memory” still functioned). He also could remember his life’s events from before his surgery, although his memories were weaker the closer the events were to the surgery. Those brain regions were clearly involved in making new long-term memories for facts or events, but not in storing old ones.

TMS

Transcranial magnetic stimulation (TMS) is a noninvasive method of creating a temporary, reversible, functional brain lesion. Using this technique, researchers disrupt the organized activity of the brain’s neurons by applying an electrical current. The current is formed by a rapidly changing magnetic field that is generated by a coil held next to the participant’s skull. The field penetrates the scalp and skull easily and causes a small current in a roughly conical portion of the

19. H.M.’s name, not publicly released until his death, was Henry Gustav Molaison. Details of his life can be found in several obituaries, including Benedict Carey, *H.M., An Unforgettable Amnesiac, Dies at 82*, N.Y. Times, Dec. 4, 2008, at A1, <https://www.nytimes.com/2008/12/05/us/05hm.html>, and H.M., *A Man Without Memories*, Economist, Dec. 18, 2008, <https://perma.cc/7CHX-UQNH>. The first scientific report of his case was William Beecher Scoville & Brenda Milner, *Loss of Recent Memory After Bilateral Hippocampal Lesions*, 20 J. Neurology Neurosurgery & Psychiatry 11 (1957), <https://doi.org/10.1136/jnnp.20.1.11>.

brain below the coil. This current induces a change in the typical responses of the neurons, which can block the normal functioning of that part of the brain.

TMS can be done in a number of ways. In some approaches, TMS happens while the participant performs the task to be studied. In these approaches, while the task is performed, TMS can be delivered as single pulses, paired pulses, or repetitive, rapid (more than once per second) pulses. Another method uses TMS for an extended period, often several minutes, before the task is performed. This sequential TMS uses slow (less than once per second) repetitive TMS.

The effects of single-pulse/paired-pulse and concurrent repetitive TMS are present while the coil is generating the magnetic field, and can extend for a few tens of milliseconds after the stimulation is turned off. By contrast, the effects of pretask repetitive TMS are thought to last for a few minutes (about half as long as the actual stimulation). When TMS is repeated regularly in nonhumans, long-term effects have been observed. Therefore, guidelines regarding how much stimulation can be applied in humans have been established.

The Food and Drug Administration (FDA) has approved TMS as a treatment for otherwise untreatable depression. The neuroscience research value of TMS stems more from its ability to create alteration of brain function in a relatively small area (about two centimeters) in an otherwise healthy brain, thus allowing for targeted testing of the role of a particular brain region for a particular class of cognitive abilities. By blocking normal functioning of the affected neurons, this can be equivalent, in effect, to a temporary lesion of that area of the brain. TMS appears to have minimal risks, but its long-term effects are not known.

DBS

Deep brain stimulation (DBS) is an FDA-approved treatment for several neurological conditions affecting movement, notably Parkinson's disease, essential tremor, and dystonia. The device used in DBS includes a lead that is implanted into a specific brain region, a pulse generator (generally implanted under the shoulder or in the abdomen), and a wire connecting the two. The pulse generator sends an electric current to the electrodes in the lead, which in turn affect the functioning of neurons in an area around the electrodes.

The precise manner by which DBS affects brain function remains unclear. Even for Parkinson's disease, for which it is widely used, individual patients sometimes benefit in unpredictable ways from placement of the lead in different locations and from different frequency or power of the stimulation.

Researchers are continuing to experiment with DBS for other conditions, such as depression, minimally conscious state, chronic pain, and overeating that leads to morbid obesity. The results are sometimes surprising. In a Canadian trial of DBS for appetite control, the obese patient did not ultimately lose weight but did suddenly develop a remarkable memory. That research group quickly started

a trial of DBS for dementia (although thus far without any obvious success).²⁰ Other surprises include some observed negative side effects from DBS, such as compulsive gambling, hypersexuality, and hallucinations. These kinds of unexpected consequences from DBS make it of continuing broader research interest.

Implanted Microelectrode Arrays

Ultimately, to understand the brain fully, one would like to know what each of its 100 billion neurons is doing at any given time, analyzed in terms of their collective patterns of activity.²¹ No current technology comes close to that kind of resolution. For example, although fMRI has a voxel size of a few cubic millimeters, it is looking at the blood flow responding to thousands or millions of neurons at each point in the brain. Conversely, while direct electrical recordings allow individual neurons to be examined and manipulated, it is not easy to record from many neurons at once. While still on a relatively small scale, recent developments now offer one method for recording from multiple neurons simultaneously by using an implanted microelectrode array.

A chip containing many tiny electrodes can be implanted directly into brain tissue. Some of those electrodes will make usable connections with neurons and can then be used either to record the activity of that neuron (when it is firing or not) or to stimulate the neuron to fire. These kinds of implants have been used in research on motor function, both in monkeys and occasionally in human patients. The research has been aimed at understanding better what neuronal activity leads to motion and hence, in the long run, perhaps to a method of treating quadriplegia or other motion disorders.

These arrays have several disadvantages as research tools. Arrays require neurosurgery for their implantation, with all of its consequent risks of infection or damage. They also have a limited lifespan, because the brain's defenses eventually prevent the electrical connection between the electrode and the neuron, usually over the span of a few months. Finally, the arrays can only reach a tiny number of the billions of neurons in the brain; current arrays have about one hundred microelectrodes. But the field and devices in use are changing rapidly. *Nature Electronics*

20. See Clement Hamani et al., *Memory Enhancement Induced by Hypothalamic/Fornix Deep Brain Stimulation*, 63 Ann. Neurology 119 (2008), <https://doi.org/10.1002/ana.21295>. See also small and mixed results reported in Andres M. Lozano et al., *A Phase II Study of Fornix Deep Brain Stimulation in Mild Alzheimer's Disease*, 54 J. Alzheimer's Disease 777 (2016), <https://pubmed.ncbi.nlm.nih.gov/27567810>.

21. See discussion in Emily R. Murphy & Henry T. Greely, *What Will Be the Limits of Neuroscience-Based Mindreading in the Law?*, in *The Oxford Handbook of Neuroethics* 635 (Judy Illes & Barbara J. Sahakian eds., 2011).

declared brain-computer interface technology the “Technology of the Year” in 2023, and dedicated a volume to its current state of development.²²

Issues in Interpreting Study Results

Lawyers trying to introduce neuroscience evidence will almost always be arguing that, when interpreted in the light of some preexisting research study, some kind of neuroscience-based test of the brain of a person in the case—usually a party, though sometimes a witness—is relevant to the case. It might be a claim that a PET scan shows that a criminal defendant was likely to have been legally insane at the time of the crime; it could be a claim that an fMRI of a witness demonstrates that they are lying. The judge will have to determine whether the scientific evidence is admissible at all under the Federal Rules of Evidence, and particularly under Rule 702. If the evidence is admissible, the finder of fact will need to consider the validity and strength of the underlying scientific finding, the accuracy of the particular test performed on the party or witness, and the application of the former to the latter.

Neuroscience-based evidence will commonly raise several scientific issues relevant to both the initial admissibility decision and the eventual determination of the weight to be given the evidence. This section will examine seven of these issues: replication, experimental design, participant selection and number, group averages, technical accuracy, statistical issues, and countermeasures. The discussion will focus on fMRI-based evidence, as that seems likely to be the method that will be used most frequently in the coming years, but most of the issues apply more broadly.

One general point is absolutely crucial. The various techniques discussed in the section titled “Some Common Neuroscience Techniques” above are generally accepted scientific procedures, for use both in research and, in most cases, in clinical care. Each one is a good scientific tool in general. The crucial issue is not likely to be whether the techniques meet the requirements for admissibility when used for *some* purposes, but whether the techniques—*when used for the purpose for which they are offered*—meet those requirements. Sometimes proponents of fMRI-based lie detection, for example, argue that the technique should be accepted because fMRI is the subject of tens of thousands of peer-reviewed publications. That is true, but irrelevant—the question is the application of fMRI to lie detection, which is the subject of far fewer, and much less definitive, publications.

22. Editorial, *An Interface Connects*, 6 Nature Elecs. 89 (2023), <https://doi.org/10.1038/s41928-023-00938-8>.

Replication

A good general rule of thumb in science is never to rely on any experimental finding until it has been independently replicated. This may be particularly true with fMRI experiments, not because of fraud or negligence on the part of the experimenters, but because, for reasons discussed below, these experiments are very complicated. Replication builds confidence that those complications have not led to false results.

In many scientific fields, including much of fMRI research, replication is sometimes not as common as it should be. A scientist often is not rewarded for replicating (or failing to replicate) another's work. Grants, tenure, and awards tend to go to people doing original research. The rise of fMRI has meant that such original experiments are easy to conceive and to attempt—any researchers with experimental expertise, access to research participants (often undergraduates), and access to an MRI scanner (found at any major medical facility) can try their own experiments and, if the study design and logic are sound and the results are statistically significant, may well end up with published results. Experiments replicating, or failing to replicate, another's work are neither as exciting nor as publishable.

For example, as discussed in more detail below, more than fifteen different laboratories have collectively published twenty to thirty peer-reviewed articles finding some statistically significant relationship between fMRI-measured brain activity and deception. None of the studies is an independent replication of another laboratory's work. Each laboratory used its own experimental design, its own scanner, and its own method of analysis. Interestingly, the published results implicate many different areas of the brain as being activated when a participant lies. A few of the brain regions are found to be important in most of the studies, but many of the other brain regions showing a correlation with deception differ from publication to publication. Only a few of the laboratories have published replications of their own work; some of those laboratories have actually published findings with different results from those in their earlier publications.

That a finding has been replicated does not mean it is correct; different laboratories can make the same mistakes. Neither does failure of replication mean that a result is wrong. Nonetheless, the existence of independent replication is important support for a finding.

Problems in Experimental Design

The most important part of an fMRI experiment is not the MRI scanner, but the design of the underlying experiment being examined in the scanner. A poorly designed experiment may yield no useful information, and even a well-designed experiment may lead to information of uncertain relevance.

A well-designed experiment must focus on the particular mental state or brain process of interest while minimizing any systematic biases. This can be especially difficult with fMRI studies. After all, these studies are measuring blood flow in the brain associated with neuronal responses in particular regions. If, for example, in an experiment trying to assess how the brain reacts to pain, the experimental participants are consistently distracted at one point in the experiment by thinking about something else, the areas of brain activation will include the areas activated by the distraction. One of the earliest published lie-detection experiments was designed so that the experimental participants pushed a button for “yes” only when saying (honestly) that they held the card displayed; they pushed the “no” button both when they did not hold the card displayed and when they did hold it but were following instructions to lie. They were to say “yes” only twenty-four times out of 432 trials.²³ The resulting differences might have come from the differences in thinking about telling the truth or telling a lie—but they also may have come from the differences in thinking about pressing the “no” button (the most common action) and pressing the “yes” button (the less frequent response). The results themselves cannot distinguish between the two explanations.

Designing good experiments is difficult, but in some respects the better the experiment, the less relevant it may prove to a real situation. A laboratory experiment attempts to minimize distractions and differences among participants, but such factors will be common in real-world settings. Perhaps more importantly, for some kinds of experiments it will be difficult, if not impossible, to reproduce in the laboratory the conditions of interest in the real world. As an extreme example, if one is interested in how a murderer’s brain functions during a murder, one cannot conduct an experiment that involves having the participant commit a murder in the scanner. For ethical reasons, that condition of interest *cannot* be tested in the experiment.

The problem of trying to detect deception provides a different example. All published laboratory-based experiments involve people who know that they are taking part in a research project. Most of them are students and are being paid to participate in the project. They have received detailed information about the experiment and have signed a consent form. Typically, they are instructed to “lie” about a particular matter. Sometimes they are told what the lie should be (to deny that they see a particular playing card, such as the seven of clubs, on a screen in the scanner); sometimes they are told to make up a lie (about their most

23. Daniel D. Langleben et al., *Telling Truth from Lie in Individual Subjects with Fast Event-Related fMRI*, 26 Hum. Brain Mapping 262 (2005), <https://doi.org/10.1002/hbm.20191>. See discussion in Nancy Kanwisher, *The Use of fMRI in Lie Detection: What Has Been Shown and What Has Not*, in Emilio Bizzi et al., *Using Imaging to Identify Deceit: Scientific and Ethical Questions* 7, 10 (2009); see also discussion in Anthony Wagner, *Can Neuroscience Identify Lies?*, in *A Judge’s Guide to Neuroscience: A Concise Introduction* 13, *supra* note 5.

recent vacation, for example). In either case, they are following instructions—doing what they *should* be doing—when they tell the “lie.”

This situation is different from the realistic use of lie detection, when a guilty person needs to tell a convincing story to avoid a high-stakes outcome, such as arrest or conviction—and even an innocent person will be genuinely nervous about the possibility of an incorrect finding of deception. In an attempt to parallel these real-world characteristics, some laboratory-based studies have tried to give participants some incentive to lie successfully; for example, the participants may be told (falsely) that they will be paid more if they “fool” the experimenters. Although this may increase the perceived stakes, it seems unlikely that it creates a realistic level of stress. These differences between the laboratory and the real world do not mean that the experimental results of laboratory studies are unquestionably different from the results that would exist in a real-world situation, but they do raise serious questions about the extent to which the experimental data bear on detecting lies in the real world.

Few judges will be expert in the difficult task of designing valid experiments. Although judges may be able themselves to identify weaknesses in experimental design, more often they will need experts to address these questions. Judges will need to pay close attention to that expert testimony and the related argument, as details of experimental design may turn out to be absolutely crucial to the value of the experimental results.

The Number and Diversity of Participants

Doing fMRI scans is expensive. The total cost of performing an hour-long research scan of a participant ranges from about \$300 to \$1,000. Much fMRI research, particularly work without substantial medical implications, is not richly funded. As a result, studies tend to use only a small number of participants—many fMRI studies use ten to twenty participants, and some use even fewer. In the lie-detection literature, for example, the number of participants used ranges from four to about thirty.

It is unclear how representative such a small group would be of the general population. This is particularly true of the many studies that use university students as research participants. Students typically are from a restricted age range, are likely to be of above-average intelligence and socioeconomic background, may not accurately reflect the country’s ethnic diversity, and typically will underrepresent people with serious mental conditions. In order to limit possible confounding variables, it can make sense for a study design to select, for example, only healthy, right-handed, native-English-speaking male undergraduates who are not using drugs. But the very process of selecting such a restricted group raises questions about whether the findings will be relevant to other groups of

people. They may be directly relevant, or they may not be. At the early stages of any fMRI research, it may not be clear what kinds of differences among participants will or will not be important.

Applying Group Averages to Individuals

Most fMRI-based research looks for statistically significant associations between particular patterns of brain activation across a number of participants. It is highly unlikely that any fMRI pattern will be found always to occur under certain circumstances in *every* person tested, or even that it will always occur under those circumstances in any one person. Human brains and their responses are too complicated for that. Research is highly unlikely to show that brain pattern “A” follows stimulus “B” each and every time and in every single person, although it may show that A follows B most of the time.

Consider an experiment with ten participants that examines how brain activation varies with the sensation of pain. A typical approach to analyzing the data is to take the average brain activation patterns of all ten participants combined, looking for the regions that, across the group, have the greatest changes—the most statistically significant changes—when the painful stimulus is applied compared to when it is absent. Importantly, though, the most significant region showing increased activation on average may not be the region with the greatest increase in activation in any particular one of the ten participants. It may not even be the area with the greatest activation in *any* of the ten participants, but it may be the region that was most consistently active across the brains of the ten participants, even if the response was small in each person.

Although group averages are appropriate for many scientific questions, the problem is that the law, for the most part, is not concerned with “average” people, but with individuals. If these “averaged” brains show a particular pattern of brain activation in fMRI studies and a defendant’s brain does not, what, if anything, does that mean?

It may or may not mean anything—or, more accurately, the chances that it is meaningful will vary. The findings will need to be converted into an assessment of an individual’s likelihood of having a particular pattern of brain activation in response to a stimulus, and that likelihood can be measured in various ways.²⁴

Consider the following simplified example. Assume that 1,000 people have been tested to see how their brain responds to a particular painful stimulus. Each is scanned twice, once when touched by a painfully hot metal rod and once when the rod is at room temperature. Assume that all of them feel pain from the heated

24. Issues of sensitivity and specificity are discussed in more detail in David H. Kaye and Hal S. Stern, *Reference Guide on Statistics and Research Methods*, in this manual.

rod and that no one feels pain from the room-temperature rod. And, finally, assume that 900 of the 1,000 show a particular pattern of brain activation when touched with the hot rod, but only 50 of the 1,000 show the same pattern when touched with the room-temperature rod.

For these 1,000 people, using the fMRI activation pattern as a test for the perception of this pain would have a sensitivity of 90% (90% of the 1,000 who felt the pain would be correctly identified and only 10% would be false negatives). Using the activation as a test for the *lack* of the pattern would have a specificity of 95% (95% of those who did not feel pain were correctly identified and only 5% were false positives). Now ask, of all those who showed a positive test result, how many were actually positive? This percentage, the positive predictive value, would be 94.7%—900 out of 950. Depending on the planned use of the test, one might care more about one of these measures than another, and there are often trade-offs between them. Making a test more sensitive (so that it misses fewer people with the sought characteristic) often means making it less specific (so that it picks up more people who do not have the characteristic in question). In any event, when more people are tested, these estimates of sensitivity, specificity, and positive predictive value become more accurate.

There are other ways of measuring the accuracy of a test of an individual, but the important point is that some such conversion is essential. A research paper that reveals that the average participant's brain (more accurately, the "averaged participants' brain") showed a particular reaction to a stimulus does not, in itself, say anything useful about how likely any one person is to have the same reaction to that stimulus. Further analyses are required to provide that information. Researchers, who are often more interested in identifying possible mechanisms of brain action than in creating diagnostic tests, will not necessarily have analyzed their data in ways that make them usefully applied to individuals—or might not have even obtained enough data for that to be possible. At least in the near future, this is likely to be a major issue for applying fMRI studies to individuals, in the courtroom or elsewhere.

Technical Accuracy and Robustness of Imaging Results

MRI machines are variable, complicated, and finicky. The machines come in several different sizes, based on the strength of the magnet, with machines used for clinical purposes ranging from 0.2T to 3.0T and research scanners going as high as 9.4T, as noted earlier. Three companies dominate the market for MRI machines—General Electric, Siemens, and Philips—although several other companies also make the machines. Both the power and the manufacturer of an MRI system can make a substantial difference in the resulting data (and images).

These variations can be more important with functional MRI (though they also apply to structural MRI), so that a result seen on a 1.5T Siemens scanner might not appear on a 3.0T General Electric machine. Similarly, results from one 3.0T General Electric machine may be different from those on an identical model.

Even the exact same MRI machine may behave differently from day to day or month to month. The machines frequently need maintenance or adjustments and sometimes can be inoperable for days or even weeks at a time. Comparing results from even the same machine before and after maintenance—or a system upgrade—can be difficult. This can make it hard to compare results across different studies or between the group average of one study and results from an individual participant.

These issues concern not only the quality of the scans done in research, but, even more importantly, the credibility of the individual scan sought to be introduced at trial. If different machines were used, care must be taken to ensure that the results are comparable. The individual scans also can have other problems. Any one scan is subject not only to machine-derived artifacts and other problems noted above, but also to human-generated artifacts, such as those caused by the participant's movements during the scan.

Finally, another technical problem of a different kind comes from the nature of fMRI research itself. The scanner will record changes in the relative levels of oxyhemoglobin to deoxyhemoglobin for thousands of voxels throughout the brain. During data analysis, these signal changes will be tested to see if they show any change in the response between the experimental condition and the baseline (or control) condition. Importantly, with fMRI, there is no definitive way to quantify precisely how large a change there was in the neural response compared to baseline; hence, the researcher must set a somewhat arbitrary statistical cutoff value (a threshold) for saying that a voxel was activated or deactivated. A researcher who wants to look only at strong effects will require a large change from baseline; a researcher who wants to see a wide range of possible effects will allow smaller changes from baseline to count.

Neither way is “right”—we do not know whether there is some minimum change in the BOLD response that means an “important” amount of brain activation has taken place, and if such a true value exists, it is likely to differ across brain regions, across tasks, and across experimental contexts. What this means is that different choices of statistical cutoff values can produce enormous differences in the apparent results. And, of course, the cutoff values used in the studies and in the scan of the individual of interest must be consistent across repeated tests. This important fact often may not be known.

Statistical Issues

Interpreting fMRI results requires the application of complicated statistical methods. These methods are particularly difficult, and sometimes controversial,

for fMRI studies, partly because of the thousands of voxels being examined. Fundamentally, most fMRI experiments look at many thousands of voxels and try to determine whether any of them are, on average, activated or deactivated as a result of the task or stimulus being studied. A simple test for statistical significance asks whether a particular result might have arisen by chance more than one time in twenty (or 5%): Is it significant at the 0.05 level? If a researcher is looking at the results for thousands of different voxels, it is likely that a number of voxels will show an effect above the threshold just by chance. There are statistical ways to control the rate of these false positives, but they need to be applied carefully. At the same time, rigid control of false positives through statistical correction (or the use of a very conservative threshold) can create another problem—an increase in the false negative rate, which results in failing to detect true brain responses that are present in the data but that fall below the statistical threshold. The community of fMRI researchers recognizes that these issues of statistical significance are difficult to resolve.

Over the past decade, other statistical techniques are increasingly being used in neuroimaging research, including techniques that do not look at the statistical significance of changes in the BOLD response in individual voxels, but that instead examine changes in the distributed patterns of activation across many voxels in a region of the brain or across the whole brain. These techniques include methods known as principal component analysis, multivariate analysis, machine learning algorithms, and artificial intelligence. These methods, the details of which are not reviewed in this reference guide, are being used increasingly in neuroimaging research and are producing some of the most interesting results in the field. The techniques are complex, and determining how to interpret the results of these tests can be controversial. Thus, these methods alone may require substantial and potentially confusing expert testimony in addition to all the other expert testimony about the underlying neuroscience evidence.

Possible Countermeasures

When neuroimaging is being used to compare the brain of one individual—a defendant, plaintiff, or witness, for example—to others, the individual undergoing neuroimaging might be able to use countermeasures to make the results unusable or misleading. And at least some of those countermeasures may prove especially hard to detect.

Individuals can disrupt almost any kind of scanning, whether done for structural or functional purposes, by moving in the scanner. Unwilling participants could ruin scans by moving their bodies or heads, or possibly even by moving their tongues. Blatant movements to disrupt the scan would be apparent, both from watching the individual in the scanner and from seeing the results, leading

to a possible negative inference that the person was trying to interfere with the scan. Nonetheless, that scan itself would be useless.

More interesting are possible countermeasures for functional scans. Polygraphy may provide a useful comparison. Countermeasures have long been tried in polygraphy with some evidence of efficacy. Polygraphy typically looks at the differences in physiological measurements of the participant when asked anxiety-provoking questions or benign control questions. Individuals can use drugs or alcohol to try to dampen their body reactions when asked anxiety-provoking questions. They can try to use mental measures to control or affect their physiological reactions, calming themselves during anxiety-provoking questions and increasing their emotional reaction to control questions. And, when asked control questions, they can try to increase the physiological signs the polygraph measures through physical means. For example, individuals might bite their tongues, step on tacks hidden in their shoes, or tighten various muscles to try to increase their blood pressure, galvanic skin response, and so on.

Basic science and polygraph research give reason for concern that polygraph test accuracy may be degraded by countermeasures, particularly when used by major security threats who have a strong incentive and sufficient resources to use them effectively. If these measures are effective, they could seriously undermine any value of polygraph security screening.²⁵

Some of the countermeasures used by a polygraph participant can be detected by, for example, drug or alcohol tests or by carefully watching the individual's body. But purely mental actions cannot be detected. These kinds of countermeasures may be especially useful to individuals seeking to beat neuroscience-based lie detection. For example, some argue that deception produces different activation patterns than telling the truth because it is mentally harder to tell a lie—more of the brain needs to work to decide whether to lie and what lie to tell. If so, two mental countermeasures immediately suggest themselves: Make the lie easier to tell (through, perhaps, memorization or practice) or make the brain work harder when telling the truth (through, perhaps, counting backward from one hundred by sevens).

Countermeasures are not, of course, potentially useful only in the context of lie detection. A neuroimaging test to determine whether a person was having the subjective feeling of pain might be fooled by the participant remembering, in great detail, past experiences of pain. The possible uses of countermeasures in neuroimaging have yet to be extensively explored, but at this point they cast additional doubt on the reliability of neuroimaging in investigations or in litigation.

25. See Nat'l Rsch. Council, *The Polygraph and Lie Detection* 5 (2003), <https://doi.org/10.17226/10420>. This report is an invaluable resource for discussions of not just the scientific evidence about the reliability of the polygraph, but also for general background about the application of science to lie detection.

Questions About the Admissibility and the Creation of Neuroscience Evidence

The admissibility of neuroscience evidence will depend on many issues, some of them arising from rules of evidence, some from the U.S. Constitution, and some from other legal provisions. Another often-overlooked reality is that judges may have to decide whether to order the creation of this kind of evidence. Certainly, judges may be called to rule on the requests for criminal defendants (or convicts seeking postconviction relief) to be able to use neuroimaging. They may also have to decide motions in civil or criminal cases to compel neuroimaging. One could even imagine requests for warrants to “search the brains” of possible witnesses for evidence, or efforts to subpoena neural data collected by third parties who manufacture or supply consumer-based wearable EEG or MEG devices. This reference guide will not seek to resolve any of these questions, but will point out some of the problems that are likely to be raised regarding admitting neuroscience evidence in court.

Evidentiary Rules

This discussion looks at the main evidentiary issues that are likely to be raised in cases involving neuroscience evidence. Note, though, that judges will not always be governed by rules of evidence. In criminal sentencing or in probation hearings, among other things, the Federal Rules of Evidence do not apply²⁶ (although federal sentencing guidelines do require that evidence used in sentencing meet a lower bar that “has sufficient indicia of reliability to support its probable accuracy”²⁷). The Rules do apply, with limitations, in other contexts.²⁸ Nonetheless, even when the Rules do not directly apply, many of the principles behind them, discussed below, will be important.

Relevance

The starting point for all evidentiary questions must be relevance. If evidence is not relevant to the questions at hand, no other evidentiary concerns matter.

26. Fed. R. Evid. 1101(d).

27. U.S. Sentencing Guidelines Manual § 6A1.3(a).

28. Fed. R. Evid. 1101(e).

This basic reminder may be particularly useful with respect to neuroscience evidence. Evidence admitted, for example, to demonstrate that a criminal defendant had suffered brain damage sometime before the alleged crime is not, in itself, relevant. The proffered fact of the defendant's brain damage must be relevant. It may be relevant, for example, to whether the defendant could have formed the necessary criminal intent, to whether the defendant should be found not guilty by reason of insanity, to whether the defendant is currently competent to stand trial, or to mitigation in sentencing. It must, however, be relevant to *something* in order to be admissible at all, and specifying its relevance will help focus the evidentiary inquiry. The question, for example, would not be whether PET scans meet the evidentiary requirements to be admitted to demonstrate brain damage, but whether they have "any tendency to make a fact more or less probable than it would be without the evidence."²⁹ The brain damage may be relevant to a fact, but that fact must be "of consequence in determining the action."³⁰

Rule 702 and the Admissibility of Scientific Evidence

Neuroscience evidence will almost always be "scientific . . . knowledge" governed by Rule 702 of the Federal Rules of Evidence, as interpreted in *Daubert v. Merrell Dow Pharmaceuticals*³¹ and its progeny, both before and after the amendments to Rule 702 in 2023.³² Rule 702 allows the testimony of a qualified expert if:

A witness who is qualified as an expert by knowledge, skill, experience, training, or education may testify in the form of an opinion or otherwise if the proponent demonstrates to the court that it is more likely than not that:

- (a) the expert's scientific, technical, or other specialized knowledge will help the trier of fact to understand the evidence or to determine a fact in issue;
- (b) the testimony is based on sufficient facts or data;
- (c) the testimony is the product of reliable principles and methods; and
- (d) the expert's opinion reflects a reliable application of the principles and methods to the facts of the case.

29. Fed. R. Evid. 401(a).

30. Fed. R. Evid. 401(b).

31. *Daubert v. Merrell Dow Pharms., Inc.*, 509 U.S. 579 (1993).

32. Rule 702 is a subject covered by Liesa L. Richter and Daniel J. Capra in *The Admissibility of Scientific Evidence*, in this manual. That reference guide should be consulted for details about Rule 702 and its application. This reference guide merely briefly lays out the Rule and its application to neuroscience evidence.

In *Daubert*, the U.S. Supreme Court listed several nonexclusive guidelines for trial court judges considering testimony under Rule 702. The committee that proposed the 2000 amendments to Rule 702 summarized these factors as follows:

The specific factors explicated by the *Daubert* Court are (1) whether the expert's technique or theory can be or has been tested—that is, whether the expert's theory can be challenged in some objective sense, or whether it is instead simply a subjective, conclusory approach that cannot reasonably be assessed for reliability; (2) whether the technique or theory has been subject to peer review and publication; (3) the known or potential rate of error of the technique or theory when applied; (4) the existence and maintenance of standards and controls; and (5) whether the technique or theory has been generally accepted in the scientific community.³³

The 2023 amendments were similarly summarized by the advisory committee as follows:

First, the rule has been amended to clarify and emphasize that expert testimony may not be admitted unless the proponent demonstrates to the court that it is more likely than not that the proffered testimony meets the admissibility requirements set forth in the rule. . . . Rule 702(d) has also been amended to emphasize that each expert opinion must stay within the bounds of what can be concluded from a reliable application of the expert's basis and methodology.³⁴

The tests laid out in *Daubert* and in the evidentiary rules governing expert testimony have been the subjects of enormous discussion, both by commentators and by courts. And, to the extent that some neuroscience evidence has been admitted in federal courts (and the courts of states that follow Rule 702 or *Daubert*), it has passed those tests. We do not have the knowledge needed to analyze them in detail, but we will merely point out a few aspects that seem especially relevant to neuroscience evidence.

Neuroscience evidence should often be subject to tests, as long as the point of the neuroscience evidence is kept in mind. An fMRI scan might provide evidence that someone was having auditory hallucinations, but it could not prove that someone was not guilty by reason of insanity. The latter is a legal conclusion, not a scientific finding. The evidence might be relevant to the question of insanity, but one cannot plausibly conduct a scientific test of whether a particular pattern of brain activation is always associated with legal insanity. One might offer neuroimaging evidence about whether a person is likely to have unusual difficulty controlling his or her impulses, but that is not, in itself, proof that the person acted recklessly. The idea of testing helps separate the conclusions that

33. Fed. R. Evid. 702 advisory committee's notes.

34. *Id.*

neuroscience might be able to reach from the legal conclusions that will be beyond it.

Daubert's stress on the presence of peer review and publication corresponds nicely to scientists' perceptions. In many scientific fields, findings that are not published in a peer-reviewed journal are deeply discounted. In those fields, scientists usually begin to have confidence in findings only after peers, both those involved in the editorial process and, more importantly, those who read the publication, have had a chance to dissect them and to search intensively for errors either in theory or in practice. It is crucial, however, to recognize that publication and peer review are not in themselves enough. The publications need to be compared carefully to the evidence that is proffered.

First, the published, peer-reviewed articles must establish the specific scientific fact being offered. An (accurate) assertion that fMRI has been the basis of more than 12,000 peer-reviewed publications will help establish that fMRI can be used in ways that the scientific community finds reliable. By themselves, however, those publications do not establish any particular use of fMRI. If fMRI is being offered as proof of deception, the twenty or thirty peer-reviewed articles concerning its ability to detect deception are most important, not the 11,980 articles involving fMRI for other purposes.

Second, the existence of several peer-reviewed publications on the same general method does not support the accuracy of any one approach if those publications are mutually inconsistent. There are now about twenty to thirty peer-reviewed publications that, using fMRI, find statistically significant differences in patterns of brain activation depending on whether the participants were telling the truth or (typically) telling a lie when instructed to do so. Many of those publications find patterns that are different from, and often inconsistent with, the patterns described in the other publications. Multiple *inconsistent* publications do not add weight to a scientific method or theory; indeed, they may subtract from it.

Third, the peer-reviewed publication needs to describe in detail the method about which the expert plans to testify. A commercial firm might, for example, claim that its method is "based on" some peer-reviewed publications, but unless the details of the firm's methods were included in the publication, those details were neither published nor peer-reviewed. A proprietary algorithm used to generate a finding published in the peer-reviewed literature is not adequately supported by that literature.

The error rate is also crucial to most neuroscience evidence, in two different senses. One is the degree to which the machines used to produce the evidence make errors. While these kinds of errors may balance out in a large sample used in published literature, any scan of any one individual may well be affected by errors in the scanning process. Second, and more importantly, neuroscience evidence will almost never give an absolute answer, but will give a probabilistic one. For example, a certain brain structure or activation pattern will be found in some percentage of people with a particular mental condition or state. These

group averages will have error rates when they are applied to individuals. Those rates need to be known and presented.

The issue of standards and controls also is important in neuroscience. This area is new and has not undergone the kind of standardization seen, for example, in forensic DNA analysis. When trying to apply neuroscience findings to an individual, evidence from the individual needs to have been acquired in the same way, with the same standards and conditions, as the evidence from which the scientific conclusions were drawn—or, at least, in ways that can be made readily comparable. For example, there is no one standard in fMRI research for what statistical threshold should be used for a change in the BOLD signal to “count” as a meaningful activation or deactivation. An individual’s scan would need to be analyzed under the same definition for activation as was used in the research supporting the method, and the effects of the chosen threshold on finding a false positive or false negative must be considered.

The final consideration—general acceptance in the scientific community—also needs to be applied carefully. There is clearly general acceptance in the scientific community that fMRI can provide scientifically and sometimes clinically useful information about the workings of human brains, but that does not mean there is general acceptance of any particular fMRI application. Similarly, there may be general acceptance that fMRI can provide some general information about the physical correlates of a particular mental state, but without general acceptance that it can do so reliably in an individual case.

Rule 403

Rule 702 is not the only test that neuroscience evidence will need to pass to be admitted in court. Even evidence admissible under that rule must still escape the exclusion provided by Rule 403: Although relevant, evidence may be excluded “if its probative value is substantially outweighed by a danger of . . . unfair prejudice, confusing the issues, misleading the jury, undue delay, wasting time, or needlessly presenting cumulative evidence.”³⁵

As discussed in detail in one article,³⁶ Rule 403 may be particularly important with some attempted applications of neuroscience evidence because of the balance it requires between the value of evidence to the decision-maker and its costs.

The probative value of such evidence may often be questioned. Neuroscience evidence will rarely, if ever, be definitive. It is likely to have a range of

35. Fed. R. Evid. 403.

36. Teneille Brown & Emily Murphy, *Through a Scanner Darkly: Functional Neuroimaging as Evidence of a Criminal Defendant’s Past Mental States*, 62 Stan. L. Rev. 1119 (2010), <https://perma.cc/C2BL-VWHV>.

uncertainties, from the effectiveness of the method in general, to questions of its proper application in this case, to whether any given individual's reactions are the same as those previously tested.

The other side of Rule 403, however, is even more troublesome. The time necessary to introduce such evidence, and to educate the jury (and judge) about it, will usually be extensive. The possibilities for confusion are likely to be great. And there is at least some evidence that jurors (or, to be precise, “mock jurors”) are particularly likely to overestimate the power of neuroscience evidence.³⁷ A high-tech “picture” of a living brain, complete with brain regions shown in bright orange and deep purple (colors not seen in an actual brain), may have an unjustified appeal to a jury. In each case, judges will need to weigh possibilities of confusion or prejudice, along with the near certainty of lengthy testimony, against the claimed probative value of the evidence.

Other Potentially Relevant Evidentiary Issues

Neuroscience evidence will, of course, be subject in individual cases to all evidentiary rules (in the Federal Rules of Evidence or otherwise) and could be affected by many of them. Four examples follow where the application of several such rules to this kind of evidence may raise interesting issues; there are undoubtedly many others.

First, in June 2009 the U.S. Supreme Court decided *Melendez-Diaz v. Massachusetts*, where the five-justice majority held that the Confrontation Clause required the prosecution to present the testimony at trial of state laboratory analysts who had identified a substance as cocaine.³⁸ This would seem to apply to any use by the prosecution in criminal cases on neuropsychological or neuroimaging of a criminal defendant or witness, although it is unclear to whom it would apply. Would testimony be required from the person who observed the procedure, the person who analyzed the results of the procedure, an AI system that analyzed the results, or programmers of that AI? These are not questions unique to neuroscience evidence, but are provoked by *Melendez-Diaz*.

37. See Deena Skolnick Weisberg et al., *The Seductive Allure of Neuroscience Explanations*, 20 J. Cognitive Neuroscience 470 (2008), <https://doi.org/10.1162/jocn.2008.20040>; David P. McCabe & Alan D. Castel, *Seeing Is Believing: The Effect of Brain Images on Judgments of Scientific Reasoning*, 107 Cognition 343 (2008), <https://doi.org/10.1016/j.cognition.2007.07.017>. These articles are discussed in Brown & Murphy, *supra* note 36, at 1199–1202. *But see* Nicholas J. Schweitzer et al., *Neuroimages As Evidence in a Mens Rea Defense: No Impact*, 17 Psych. Pub. Pol’y & L. 357 (2011), <https://doi.org/10.1037/a0023581> (explaining that the experimental results seem to indicate that showing neuroimages to mock jurors does not affect their decisions).

38. 557 U.S. 305 (2009).

Second, the Federal Rules of Evidence put special limits on the admissibility of evidence of character and, in some cases, of predisposition.³⁹ In some cases, neuroscience evidence offered for the purpose of establishing a regular behavior of the person might be viewed as evidence of character⁴⁰ or predisposition (or lack of predisposition). Whether such evidence could be admitted might hinge on whether it was offered in a civil case or a criminal case, and, if in a criminal case, by the prosecution or the defendant.

Third, Federal Rule of Evidence 406 allows the admission of evidence about a habit or routine practice to prove that the relevant person's actions conformed to that habit or routine practice. It is conceivable that neuroscience evidence might be used to describe "habits of mind" and thus be offered under this rule.

The fourth example applies to neuroscience-based lie detection. Although New Mexico is the only U.S. jurisdiction that generally allows the introduction of polygraph evidence,⁴¹ several jurisdictions allow polygraph evidence in two specific situations. First, polygraph evidence is sometimes allowed when both parties have stipulated to its admission in advance of the performance of the test. (This does lead one to wonder whether a court would allow evidence from a psychic or from a fortune-telling toy, like the Magic 8 Ball, if both parties stipulated to it.) Second, polygraph evidence is sometimes allowed to impeach or to

39. Fed. R. Evid. 404, 405, 412–15, 608.

40. Evidence about lie detection has sometimes been viewed as "character evidence," introduced to bolster a witness's credibility. The Canadian Supreme Court has held that polygraph evidence is inadmissible in part because it violates the rule limiting character evidence:

What is the consequence of this rule in relation to polygraph evidence? Where such evidence is sought to be introduced, it is the operator who would be called as the witness, and it is clear, of course, that the purpose of his evidence would be to bolster the credibility of the accused and, in effect, to show him to be of good character by inviting the inference that he did not lie during the test. In other words, it is evidence not of general reputation but of a specific incident, and its admission would be precluded under the rule. It would follow, then, that the introduction of evidence of the polygraph test would violate the character evidence rule.

R. v. Béland, [1987] 2 S.C.R. 398, 414. The Canadian court also held that polygraph evidence violated another rule concerning character evidence, the rule against "oath-helping":

From the foregoing comments, it will be seen that the rule against oath-helping, that is, adducing evidence solely for the purpose of bolstering a witness's credibility, is well grounded in authority. It is apparent that, since the evidence of the polygraph examination has no other purpose, its admission would offend the well-established rule.

Id. at 408 (the court also ruled against polygraph evidence as violating the rule against prior consistent statements and, because the jury needs no help in assessing credibility, the rule on the use of expert witnesses.).

41. *Lee v. Martinez*, 96 P.3d 291 (N.M. 2004).

corroborate a witness's testimony.⁴² If a neuroscience-based lie-detection technique were found to be as reliable as the polygraph, presumably those jurisdictions would have to consider whether to extend these exceptions to such neuroscience evidence.

Constitutional and Other Substantive Rules

In many contexts, courts will be asked to admit neuroscience evidence or to order, allow, or punish its creation. Such actions may implicate a number of constitutional rights, international law, and international human rights, as well as other substantive legal and regulatory provisions. While this section will not seek to discuss all possible such claims or to resolve any of them, it raises some of the most interesting issues.

Possible Rights Against Neuroscience Evidence

Much of the discussion that follows presumes the compulsory creation of neuroscience data, or compulsory submission to neuroimaging or other testing that would generate neuroscience data. As brain sensors become more widely integrated into everyday headphones, earbuds, and even wrist-worn devices to detect motor neuron activity, neuroscience data may be increasingly created passively, without compulsion, which would reshape the legal analysis that would follow. Passively created data may be subject to discovery through subpoenas to third parties who operate the devices or consumer applications that collect the data generated by these devices.

The Fifth Amendment privilege against self-incrimination

Could a person be required to submit to neuroimaging to gather neuroscience data from their brains, or would that violate their privilege against self-incrimination? This has been discussed by legal scholars for more than a decade now because of the novel fact that neuroscience evidence could be seen as “real” physical evidence rather than “testimonial” evidence because it measures brain activity and blood flow rather than spoken words by an individual. The real-world use in the United States of such evidence has been quite limited, so the question remains a theoretical one. For what purpose might a person be asked to

42. See, e.g., *United States v. Piccinonna*, 885 F.2d 1529 (11th Cir. 1989) (en banc). See also *United States v. Allard*, 464 F.3d 529 (5th Cir. 2006); *Thornburg v. Mullin*, 422 F.3d 1113 (10th Cir. 2005).

submit to such examination? To gather identifying evidence, automatically processed memories, or silently uttered responses in their brain.⁴³

If neuroscience evidence were held to be “testimonial” and its creation compelled rather than “real” physical evidence passively created, it would be subject to the privilege, but if it were nontestimonial and not compelled, it would, under current law, not be. Examples of nontestimonial evidence for purposes of the privilege against self-incrimination include a blood alcohol test or medical X-rays. Examples of evidence created without compulsion include a private diary or other business documents created without government instruction to do so.

It is not the *type* of neuroimaging that addresses whether the privilege against self-incrimination is implicated but the nature of the evidence collected, the inferences that can be drawn from it, and whether a person is compelled to create it. How the evidence from the brain is generated (passively or in response to questions, for example) and the nature of the neuroscience evidence sought (evidence of brain damage, for example, versus silently uttered responses to questions or recalled memories that could infer unspoken thoughts) will be more relevant to deciding whether it runs afoul of the privilege against self-incrimination.

It is possible, however, that conscious answers may not be necessary but the brain could be queried for memories or other information nonetheless. Two EEG-based systems claim to be able to determine whether a person either recognizes or has “experiential knowledge” of an event (a memory derived from experience as opposed to being told about it).⁴⁴ Very substantial scientific questions exist about each system, but, assuming they were to be admitted as reliable, they would raise this question more starkly because they do not require the participant in the procedure to consciously create answers or respond to questions. The participant is shown photographs of relevant locations or read a description of the events while hooked up to an EEG. The brain waves, it is

43. See discussion of “spectrum of evidence” in Nita A. Farahany, *Incriminating Thoughts*, 64 Stan. L. Rev. 351 (2012), <https://perma.cc/2VX7-N6P7>.

44. The first system is the so-called Brain Fingerprinting, developed by Dr. Larry Farwell. This method was introduced successfully in evidence at the trial court level in a postconviction relief case in Iowa; the use of the method in that case is discussed briefly in the Iowa Supreme Court’s decision on appeal. *Harrington v. State*, 659 N.W.2d 509, 516 n.6 (Iowa 2003). The court expressed no view on whether that evidence was properly admitted. See *id.* at 516. The method is discussed on the website of Farwell’s company, Brain Fingerprinting Laboratories. It is discussed from a scientific perspective in J. Peter Rosenfeld, ‘Brain Fingerprinting’: A Critical Analysis, 4 Sci. Rev. Mental Health Prac. 20 (2005), <https://perma.cc/R4KF-92AS>. See also the brief discussion in Henry T. Greely & Judy Illes, *Neuroscience-Based Lie Detection: The Urgent Need for Regulation*, 33 Am. J.L. & Med. 377, 387–88 (2007), <https://doi.org/10.1177/009885880703300211>.

The second system is called Brain Electrical Oscillation Signature (BEOS) and was developed in India, where it has been introduced in trials and has been important in securing criminal convictions. See Anand Giridharadas, *India’s Novel Use of Brain Scans in Courts Is Debated*, N.Y. Times, Sep. 14, 2008, at A10, <https://www.nytimes.com/2008/09/15/world/asia/15brainscan.html>.

asserted, demonstrate whether the individual recognizes the photographs or has “experiential knowledge” of the events—no volitional communication is necessary. It might be harder to classify these automatically processed responses as “testimonial,” compared to unspoken silent utterances in response to questions.⁴⁵

Other possible legal protections against compulsory neuroscience procedures

Even if the privilege against self-incrimination applies to neuroscience methods of obtaining evidence, it only applies where someone invokes the privilege. The courts and other government bodies force people to answer questions all the time, often under penalty of criminal or civil sanctions or of the court’s contempt power. And the government can immunize a person from criminal prosecution and compel responses without running afoul of the privilege against self-incrimination. For example, a plaintiff in a civil case alleging damage to their health can be compelled to undergo medical testing at a defendant’s appropriate request. In that case, the plaintiff can refuse, but only at the risk of seeing their case dismissed. Presumably, a party could similarly demand that a party, or a witness, undergo a neuropsychological or neuroimaging examination, looking for either structural or functional aspects of the person’s brain or mental status relevant to the case or, for example, relevant to the accuracy of their memories.⁴⁶ If the privilege against self-incrimination is not available, or is available but not attractive, could the person asked have any other protection?

The answer is not clear. One might try to argue, along the lines of *Rochin v. California*,⁴⁷ that such a procedure violates the Due Process Clause of the Fifth and Fourteenth Amendments because it intrudes on the person in a manner that “shocks the conscience.” One might try to use the concept of “freedom of thought” referenced in some U.S. Supreme Court First Amendment cases to argue that the First Amendment’s freedoms of religion, speech, and the press encompass a broader protection of the contents of the mind through a right to cognitive liberty.⁴⁸ Until recently, governments could not directly access our thoughts⁴⁹—which is why freedom of thought as a fundamental right has largely been glossed over in history. Justices Murphy, Black, and Douglas, along with Chief Justice Stone, reflected this prevailing view in their 1942 dissenting and

45. Farahany, *supra* note 43; Farahany, *supra* note 2.

46. See, e.g., *Stalcup v. State*, 311 P.3d 104 (Wyo. 2013).

47. *Rochin v. California*, 342 U.S. 165 (1952).

48. See Paul Root Wolpe, *Is My Mind Mine? Neuroethics and Brain Imaging*, in *The Penn Center Guide to Bioethics* 86 (Arthur L. Caplan, Autumn Fiester & Vardit Ravitsky eds., 2009).

49. Simon McCarthy-Jones, *The Autonomous Mind: The Right to Freedom of Thought in the Twenty-First Century*, 2 *Frontiers A.I.* 1 (2019), <https://doi.org/10.3389/frai.2019.00019>.

concurring opinion in the U.S. Supreme Court case of *Jones v. City of Opelika*, when they argued that while freedom of thought is absolute, even “the most tyrannical government is powerless to control the inward workings of the mind.”⁵⁰ Now that there are ways to infer thought using neurotechnology, we find ourselves with very little history or theoretical work to help define its scope.

Scholars have advanced a theory of “cognitive liberty” as a human right that would include the right to mental privacy as part of the human right to privacy under the Universal Declaration of Human Rights (UDHR), the right to keep one’s thoughts private and not be punished for one’s thoughts under the right to freedom of thought in the UDHR, and a right to self-determination over one’s brain and mental experiences.⁵¹

That right, as one of this reference guide’s authors has argued, has been implicitly recognized in tort cases in the United States, which do not extend a duty to mitigate one’s injuries to measures that would tamper with memories or reduce one’s psychological injuries in the same manner that other physical injuries must be mitigated or one’s damages will be reduced in civil cases.⁵² And it is implicitly recognized in First Amendment cases, where freedom of thought is recognized as a precursor to freedom of speech.

But all of these rights and their applications and implications for neuroscience evidence remain speculative until courts grapple with new forms of neuroscience evidence and challenges to their creation or introduction in legal cases.

Other substantive rights against neuroscience evidence

At least one form of possible neuroscience evidence may already be covered by statutory provisions limiting its creation and use—lie detection. In 1988, Congress passed the federal Employee Polygraph Protection Act (EPPA).⁵³ Under this Act, almost all employers are forbidden to “directly or indirectly . . . require, request, suggest, or cause any employee or prospective employee to take or submit to any lie detector test” or to “use, accept, refer to, or inquire concerning the results of any lie detector test of any employee or prospective employee.”⁵⁴ The Act defines a “lie detector” broadly, as “a polygraph, deceptiongraph, voice stress

50. 316 U.S. 584, 618 (1942) (Murphy, J., dissenting).

51. See, e.g., Farahany, *supra* note 2.

52. See, e.g., Nita A. Farahany, *The Costs of Changing Our Minds*, 69 Emory L.J. 75 (2019), <https://perma.cc/L488-XMBW>.

53. Federal Employee Policy Protection Act of 1988, Pub. L. No. 100–347, 102 Stat. 646 (codified at 29 U.S.C. §§ 2001–2009). See generally the discussion of federal and state laws in Greely & Illes, *supra* note 44, at 405–10, 421–31.

54. 29 U.S.C. § 2002 (1)–(2) (The section also prohibits employers from taking action against employees because of their refusal to take a test, because of the results of such a test, or for asserting their rights under the Act.); see *id.* § 2001 (3)–(4) (definitions).

analyzer, psychological stress evaluator, or any other similar device (whether mechanical or electrical) that is used, or the results of which are used, for the purpose of rendering a diagnostic opinion regarding the honesty or dishonesty of an individual.”⁵⁵ The Department of Labor can punish violators with civil fines, and those injured have a private right of action for damages.⁵⁶ The Act does provide narrow exceptions for polygraph tests in some circumstances.⁵⁷

In addition to federal statutes, many states passed their own versions of the EPPA, either before or after the federal act. The laws passed after the EPPA generally apply similar prohibitions to some employers not covered by the federal act (such as state and local governments), but with their own idiosyncratic set of exceptions. Many states have also passed laws regulating lie-detection services. Most of these seem clearly aimed at polygraphy, but, in some states, the language used is quite broad and may well encompass neuroscience-based lie detection.⁵⁸

States also may provide protection against neuroscience evidence that goes beyond lie detection and could prevent involuntary neuroscience procedures. Some states have constitutional or statutory rights of privacy that could be read to include a broad freedom for mental privacy. And in some states, such as California, such privacy rights apply not just to state action but to private actors as well.⁵⁹ Most employment cases would be covered by the EPPA and its state equivalents, but such state privacy protections might be used to help decide whether courts could compel neuroimaging scans or whether they could be required in nonemployment relationships, such as school/student or parent/child.

Neuroscience evidence and the Sixth and Seventh Amendment rights to trial by jury

One might also argue that some kinds of neuroscience evidence could be excluded from evidence as a result of the federal constitutional rights to trial by

55. *Id.* § 2001(3).

56. *Id.* § 2005.

57. *Id.* § 2006.

58. *See generally* Greely & Illes, *supra* note 44, at 409–10, 421–31 (for both state laws on employee protection and state laws more broadly regulating polygraphy).

59. “All people are by nature free and independent and have inalienable rights. Among these are enjoying and defending life and liberty, acquiring, possessing, and protecting property, and pursuing and obtaining safety, happiness, and privacy.” Calif. Const., art. I, § 1 (emphasis added). The words “and privacy” were added by constitutional amendment in 1972. The California Supreme Court had applied these privacy protections in suits against private actors: “In summary, the Privacy Initiative in article I, section 1 of the California Constitution creates a right of action against private as well as government entities.” *Hill v. Nat’l Collegiate Athletic Ass’n*, 865 P.2d 633, 644 (Cal. 1994).

jury in criminal and most civil cases. In *United States v. Scheffer*,⁶⁰ the U.S. Supreme Court upheld an express ban in the Military Rules of Evidence on the admission of any polygraph evidence against a criminal defendant's claimed Sixth Amendment right to introduce the evidence in their defense. Justice Thomas wrote the opinion of the Court, holding that the ban was justified by the questionable reliability of the polygraph. Justice Thomas continued, however, in a portion of the opinion joined only by Chief Justice Rehnquist and Justices Scalia and Souter, to hold that the Rule could also be justified by an interest in the role of the jury:

It is equally clear that Rule 707 serves a second legitimate governmental interest: Preserving the jury's core function of making credibility determinations in criminal trials. A fundamental premise of our criminal trial system is that "the jury is the lie detector." Determining the weight and credibility of witness testimony, therefore, has long been held to be the "part of every case [that] belongs to the jury, who are presumed to be fitted for it by their natural intelligence and their practical knowledge of men and the ways of men."⁶¹

The other four justices in the majority, and Justice Stevens in dissent, disagreed that the role of the jury justified this rule, but the question remains open. Justice Thomas's opinion did not argue that exclusion was *required* as part of the rights to jury trials in criminal and civil cases under the Sixth and Seventh Amendments, respectively, but one might try to extend his statements of the importance of the jury as "the lie detector" to such an argument.⁶²

Possible Rights to the Creation or Use of Neuroscience Evidence

The Eighth Amendment right to present evidence of mitigating circumstances in capital cases

In one of many ways in which "death is different," in *Lockett v. Ohio*,⁶³ the U.S. Supreme Court held that the Eighth Amendment guarantees a convicted defendant in a capital case a sentencing hearing in which the sentencing authority is able to consider any mitigating factors. In *Rupe v. Wood*,⁶⁴ the Ninth Circuit, in

60. 523 U.S. 303 (1998).

61. *Id.* at 312–13 (internal citation omitted).

62. The Federal Rules of Criminal Procedure effectively give the prosecution a right to a jury trial, by allowing a criminal defendant to waive such a trial only with the permission of both the prosecution and the court. Fed. R. Crim. P. 23(a). Many states allow a criminal defendant to waive a jury trial unilaterally, thus depriving the prosecution of an effective "right" to a jury.

63. 438 U.S. 586 (1978).

64. 93 F.3d 1434, 1439–41 (9th Cir. 1996). *But see* *United States v. Fulks*, 454 F.3d 410, 434 (4th Cir. 2006). *See generally* Christopher Domin, Comment, *Mitigating Evidence? The Admissibility of*

an appeal from the defendant's successful habeas corpus proceeding, applied that holding to find that a capital defendant had a constitutional right to have polygraph evidence admitted as mitigating evidence in his sentencing hearing. The court agreed that totally unreliable evidence, such as astrology, would not be admissible, but that the district court had properly ruled that polygraph evidence was not that unreliable. (The Washington Supreme Court had previously decided that polygraph evidence should be admitted in the penalty phase of capital cases under some circumstances.⁶⁵) Thus, capital defendants may argue that they have the right to present neuroscience evidence as mitigation even if it would not be admissible during the guilt phase.

The Sixth Amendment right to present a defense

The *Scheffer* case arose in the context of another right guaranteed by the Sixth Amendment, the right of a criminal defendant to present a defense. It seems likely that neuroimaging will first be offered by parties who have been its voluntary participants and who will argue that it strengthens their cases, just as other kinds of neuroscience evidence have. In fact, the main use of neuroimaging in the courts so far, at least in criminal cases, has been by defendants seeking to demonstrate a defense or mitigation of sentencing. If jurisdictions were to exclude such evidence categorically, they might face a similar Sixth Amendment challenge.

The U.S. Supreme Court has held that some prohibitions on evidence in criminal cases violate the right to present a defense. Thus, in *Rock v. Arkansas*,⁶⁶ the Court struck down a per se rule in Arkansas against the admission of hypnotically refreshed testimony, holding that it was “arbitrary or disproportionate to the purposes [it is] designed to serve.” The *Scheffer* case probably provides the model for how arguments about exclusions of neuroscience evidence would play out. Eight of the Justices in *Scheffer* agreed that the reliability of polygraphy was sufficiently questionable as to justify the per se ban on its use. Justice Stevens, however, dissented, finding polygraphy sufficiently reliable to invalidate its per se exclusion.

The Fourth Amendment

The Fourth Amendment raises some particularly interesting questions. It provides, of course, that, “The right of the people to be secure in their persons,

Polygraph Results in the Penalty Phase of a Capital Trial, 43 U.C. Davis L. Rev. 1461 (2010), <https://perma.cc/JA6C-DLGH> (arguing that the Supreme Court should resolve the resulting circuit split by adopting the Ninth Circuit's position).

65. *State v. Bartholomew*, 683 P.2d 1079, 1083–84 (Wash. 1984).

66. 483 U.S. 44, 56 (1987).

houses, papers, and effects, against unreasonable searches and seizures, shall not be violated, and no Warrants shall issue, but upon probable cause, supported by Oath or affirmation, and particularly describing the place to be searched, and the persons or things to be seized.” On the one hand, an involuntary neuroscience examination would seem to be a search or seizure and thus “unreasonable” neuroscience examinations are prohibited. To that extent, the Fourth Amendment would appear to be a protection against compulsory neuroscience testing.

On the other hand, neuroimaging may be considered noninvasive in that one does not have to physically invade a person’s body or effects to obtain information from the brain. A suspect can be compelled to undergo compulsory urine or blood tests, for example, and might similarly be compelled to undergo neuroimaging without it being considered an “unreasonable” search under the Fourth Amendment. Not all evidence obtained from the brain will be equally sensitive—one could obtain identifying information, automatic functioning information such as the processing of alcohol, memories from the brain, or silent utterances. Courts will be confronted with having to parse not just whether a search has occurred, but whether that search is unreasonable when neuroscience evidence is compelled. And the answers will likely turn on the nature of the neuroimaging used, whether it is compelled or passively collected neuroscience information from a wearable device, and the nature of the evidence being sought and the privacy interest implicated as a result.⁶⁷

Examples of the Possible Uses of Neuroscience in the Courts

Most of what follows deals with criminal cases. That is the area where the interest in neuroscience evidence has been strongest, and the scholarly literature, in both neuroscience and in law, has largely reflected that focus. But it is important to recognize that in U.S. civil cases, neuroscience evidence has been used to challenge or support the veracity of eyewitness memory, the reality and intensity of pain, and the mental capacity and competence of parties, witnesses, and testators.

For example, neuroscience has increasingly played an evidentiary role in civil cases involving allegations of brain injury, such as traumatic brain injury (TBI), stroke, or dementia.⁶⁸ In a product liability case involving a tractor

67. See, e.g., Nita A. Farahany, *Searching Secrets*, 160 U. Pa. L. Rev. 1239 (2012), <https://perma.cc/DU22-SDBU>.

68. See, e.g., *Dickson v. United Airlines*, No. 4:20-CV-5014-RMP, 2021 WL 4199954 (E.D. Wash. Aug. 3, 2021) (dispute between experts on neuroimaging findings and tier consistency with TBI).

accident that resulted in the plaintiff's facial and traumatic brain injury, neuroscience evidence played a central role. The plaintiff's expert testimony, partially based on magnetic resonance imaging (MRI) and diffusion tensor imaging (DTI), was challenged by the defendant. However, the court upheld the admissibility of this evidence under Rule 702 of the Federal Rules of Evidence, recognizing DTI as a reliable methodology for identifying TBI.⁶⁹ Other decisions that have excluded DTI as scientifically unreliable have been challenged by scholars in the field.⁷⁰

Neuroscience has also been used in civil cases involving claims of chronic pain—such as fibromyalgia, complex regional pain syndrome (CRPS), or phantom limb pain—to challenge or corroborate the subjective reports of pain by the plaintiffs, including in claims and challenges for long-term disability payments.⁷¹ However, a consensus statement by neuroscientists, ethicists, and legal scholars in *Nature Reviews Neurology* opined that neuroscience evidence of chronic pain should be used to educate decision-makers about the underlying pathophysiology of the pain, but that “brain imaging is not yet sufficiently reliable to be used as a pain detector to either support or contradict an individual's self-report.”⁷²

Civil cases involving issues of memory, such as false or recovered memories, eyewitness identification, or consent, have also increasingly involved neuroscientific studies, often to explain or challenge the accuracy and reliability of witness or party testimony. However, this trend has raised scholarly concerns that neuroimaging can disproportionately influence juror assessments, particularly in determining awards in cases involving claims of pain and suffering.⁷³ All of these civil uses, and more, are likely to expand in coming years.

But we have some data about criminal cases. Over the past two decades, neuroscience evidence has increasingly been introduced by criminal defendants to challenge their competency and mental state, to put into question the voluntariness of their conduct, to support claims of insanity and mental diseases or defects, and to mitigate their punishment. Neuroscience evidence has been used to challenge the memory of eyewitnesses. And it could increasingly be used to challenge or interrogate the memory of witnesses, criminal suspects, or litigants

69. *White v. Deere & Co.*, No. 13-CV-02173-PAB-NYW, 2016 WL 462960 (D. Colo. Feb. 8, 2016).

70. See, e.g., van Velkinburgh, Herbst & Casper, *supra* note 15.

71. See, e.g., *Gilbert v. Principal Life Ins. Co.*, No. 8-CV-TDC-21-0128, 2022 WL 3369537 (D. Md. Aug. 16, 2022).

72. Karen D. Davis et al., *Brain Imaging Tests for Chronic Pain: Medical, Legal and Ethical Issues and Recommendations*, 13 *Nature Revs. Neurology* 624, 635 (2017), <https://doi.org/10.1038/nrneurol.2017.122>.

73. See generally Hannah J. Phalen, Jessica M. Salerno & Nicholas J. Schweitzer, *Can Neuroimaging Prove Pain and Suffering?: The Influence of Pain Assessment Techniques on Legal Judgments of Physical Versus Emotional Pain*, 45 *L. & Hum. Behav.* 393 (2021), <https://doi.org/10.1037/lhb0000460>.

in civil and criminal cases. There are very few cases, civil or criminal, where the mental states of the parties are not at least theoretically relevant on issues of competency, intent, motive, recklessness, negligence, good or bad faith, memory, or other issues. And even if the parties' own mental states were not relevant, the mental states of witnesses almost always will be potentially relevant—are they telling the truth? Are they biased against one party or another? Is their memory as reliable as they claim it to be? The mental states and memory of jurors, and even of judges, occasionally may be called into question.⁷⁴

There are some important caveats on understanding the use of neuroscience in legal proceedings. First, while neuroscience evidence has been increasingly used, the evidence is usually based on neuropsychological testing and not neuroimaging. In criminal cases, neuroscience evidence has been most often introduced when the defendant is charged with a serious felony offense, rather than more minor misdemeanors, because it is expensive to investigate the neurological status of the defendant and thus impractical to introduce in misdemeanor cases. The garden-variety breach of contract or assault and battery is also not likely to provide a plausible context for convincing neuroscience evidence, especially if there is no evidence that the actor or the actions were odd or bizarre. And many cases will not provide, or justify, the resources necessary for a neurological evaluation or neuroimaging.

Second, even when neuroscience evidence is introduced, it often has a “time machine” problem. The neurological examination of a criminal defendant may be taking place months or years after the offense. Neuroscience-based testing after the crime is unlikely to discern a person's state of mind in the past. Unless the legally relevant action took place inside an MRI scanner or while a defendant was using wearable brain sensors, the best it may be able to do is to say that, based on the defendant's current mental condition or state, as shown by the current brain structure or functioning, the defendant is more or less likely than average to have had a particular mental state or condition at the time of the relevant event. If the time of the relevant event is the time of trial (or shortly before trial)—as would be the case with the truthfulness of testimony, the existence of bias, or the existence of a particular memory—that would not be a problem, but otherwise it would be.

The possibility of using real-time recordings of brain activity has reportedly already been attempted. An accused sought evidence of his brain activity recorded through electrodes implanted to address epileptic seizures as evidence he was suffering from an epileptic seizure at the time of an alleged attack.⁷⁵ Fitbit

74. See Emily Murphy, *Evidence: Evidence of Memory from Brain Data*, 5 Judges' Book 67 (2021), <https://perma.cc/69VY-EENM>; Emily R.D. Murphy & Jesse Rissman, *Evidence of Memory from Brain Data*, 7 J.L. & Biosciences 1 (2020), <https://doi.org/10.1093/jlb/ljaa078>.

75. Jessica Hamzelou, *How Your Brain Data Could Be Used Against You*, MIT Tech. Rev., Feb. 24, 2023, <https://perma.cc/68PP-HYKW>.

data have similarly been used in several cases as circumstantial evidence to corroborate or disprove criminal case evidence.⁷⁶

Neuroscience evidence has been increasingly introduced into criminal cases and accepted in a number of such cases. More than 2,800 judicial opinions from 2005 to 2015 discuss the use of neuroscience evidence by a criminal defendant as part of their defense.⁷⁷ In general, this follows an evaluation of the defendant for specific neurological abnormalities that may at least in part explain, or partially excuse or mitigate, a defendant's behavior. This includes medical history, neuropsychological testing, brain scanning, or unsubstantiated claims of a past history of head trauma or brain injury. Since 2015, hundreds more cases per year have discussed similar claims.

In the majority of these criminal cases, the most serious charge against the criminal defendant is some degree of homicide, but in about 40% of the cases, the most serious charge is a felony other than homicide. In only about 10% of the cases do the judicial opinions discuss the use of neuroimaging as part of the neuroscience evidence introduced.⁷⁸

In some cases, rather than addressing the unique issues relevant to a specific person's brain, it has been used to inform broader policy choices about categories of individuals. This is particularly true for juvenile defendants and those under the age of twenty-six, who argue that their developing brain, as a class of individuals, warrants treating them differently than adults. Developmental neuroscience, which has shown that adolescents have developing brains that make conforming with the law more difficult than it is for adults, has been the empirical basis for recent constitutional prohibitions against the execution of—or punishment of life without the possibility of parole for—juveniles.⁷⁹ And an increasing number of criminal defendants have argued, albeit largely unsuccessfully, that the same logic ought to extend to them until at least twenty-six years of age, when the brain has by and large fully developed.⁸⁰ Three of the handful of published cases in which fMRI evidence was offered in court concerned challenges to state laws requiring warning labels on violent video games.⁸¹ The states

76. Farahany, *supra* note 2, at 83.

77. Henry T. Greely & Nita A. Farahany, *Neuroscience and the Criminal Justice System*, 2 Ann. Rev. Criminology 451, 453 (2019), <https://doi.org/10.1146/annurev-criminol-011518-024433>.

78. *Id.* at 454–55.

79. *Graham v. Florida*, 560 U.S. 48 (2010); *Miller v. Alabama*, 567 U.S. 460 (2012); *Roper v. Simmons*, 543 U.S. 551 (2005).

80. Francis X. Shen et al., *Justice for Emerging Adults After Jones: The Rapidly Developing Use of Neuroscience to Extend Eighth Amendment Miller Protections to Defendants Ages 18 and Older*, 97 N.Y.U. L. Rev. 101 (2022), <https://perma.cc/4GZV-R5E5>.

81. *Ent. Software Ass'n v. Hatch*, 443 F. Supp. 2d 1065 (D. Minn. 2006); *Ent. Software Ass'n v. Blagojevich*, 404 F. Supp. 2d 1051 (N.D. Ill. 2005); *Ent. Software Ass'n v. Granholm*, 404 F. Supp. 2d 978 (E.D. Mich. 2005). Each of the three courts held that the state statutes violated the First Amendment.

sought, without success, to use fMRI studies of the effects of violent video games on the brains of children playing the games to support their statutes.⁸²

These “wholesale” uses of neuroscience may (or may not) end up affecting the law, but the courts would be more affected if various “retail” uses of neuroscience were to become common, where a party or a witness is subjected to neuroscience procedures to determine something relevant only to that particular case. An incomplete list of some of the most plausible categories for such retail uses includes the following:

- issues of responsibility, certainly criminal and likely also civil
- predicting future behavior for sentencing
- mitigating (or potentially aggravating) factors on sentencing
- competency, now or in the past, to take care of one’s affairs, to enter into agreements or make wills, to stand trial, to represent oneself, and to be executed
- credibility or deception in current statements
- bolstering or as evidence of a mental disease or defect in the insanity defense
- the existence or nonexistence of a memory of some event and, possibly, some information about the status of that memory (true, false; new, old, etc.)
- the presence of the subjective sensation of pain
- the presence of bias against a party

Many, but not all, of these issues have begun to be discussed in the literature. A few of them, such as criminal responsibility, mitigation, memory detection, and lie detection, are appearing in courtrooms; others, such as pain detection, have been introduced in a number of civil cases. This reference guide does not discuss all of these topics and does not discuss any of them in great depth, but it will discuss three of them substantially—criminal responsibility, detection of pain, and lie detection—in order to provide a flavor of the possibilities, as well as presenting a short discussion of a fourth: addiction.

82. The courts, sitting in equity and so without juries, all considered the scientific evidence and concluded that it was insufficient to sustain the statutes’ constitutionality. In *Blagojevich* the court heard testimony for the state directly from Dr. Kronenberger, the author of some of the fMRI-based articles on which the state relied, as well as from Dr. Howard Nusbaum, for the plaintiffs, who attacked Dr. Kronenberger’s studies. After a substantial discussion of the scientific arguments, the district court judge, Judge Matthew Kennelly, found that “Dr. Kronenberger’s studies cannot support the weight he attempts to put on them via his conclusions,” and did not provide a basis for the statute. *Blagojevich*, 404 F. Supp. at 1063–67. Judge Kennelly’s discussion of this point may be a good example of the kind of analysis neuroscience evidence may force upon judges.

Criminal Responsibility

Neuroscience may raise some deep questions about criminal responsibility. Assume we had excellent scientific evidence that a defendant could not help but commit the criminal acts because of a specific brain abnormality.⁸³ Should that affect the defendant's guilt and, if so, how? Should it affect their sentence or other subsequent treatment? The moral questions may prove daunting. Currently, the law assumes "legal" free will rather than engaging with the "philosophical" debates over free will.⁸⁴ The criminal law in all U.S. jurisdictions presumes that individuals are responsible agents capable of making choices and intending the natural consequence of their actions. While some commentators believe that neuroscience should change that assumption in criminal law, "[a]cts should be judged by their tendency under known circumstances, not by the actual intent which accompanies them."⁸⁵ The criminal defendant may nonetheless evade or receive a diminution in criminal responsibility through a successful presentation of a justification or excuse.

A finding of criminal liability requires the government to prove the actor voluntarily engaged in a harmful or threatening act proscribed by criminal law and did so with the requisite mental awareness of the circumstances of fact that made the conduct criminal. A conviction generally requires both an *actus reus* and a *mens rea*—a "guilty act" and a "guilty mind." An unconscious person cannot "act," but even a conscious act is often not enough. Specific crimes often require specific intents, such as acting with a particular purpose or in a knowing, reckless, or negligent way. Some crimes require even more defined mental states, such as a requirement for specific intent to not only commit a homicide, but to do so with knowledge of the special status of the victim, such as a police officer. And almost all crimes can be excused by legal insanity. In these and other ways, the mental state of the defendant may be relevant to a criminal case.

Neuroscience may provide evidence in some cases to support a defendant's claim of a justification or excuse by helping to establish that they suffer a mental disease or disorder. For example, a defendant who claims to have been insane at the time of the crime might try to support their claim by alleging that they are suffering from frontotemporal dementia, which prevented them from knowing the difference between right and wrong. Neuroimaging may be able to provide some evidence about whether the defendant is, in fact, suffering from frontotemporal dementia and how that affects the individual, at least at the time they

83. See Henry T. Greely, *Neuroscience and Criminal Responsibility Proving "Can't Help Himself" as a Narrow Bar to Criminal Liability*, in 13 *Law & Neuroscience*, Current Legal Issues 61 (Michael Freeman ed., 2011), <https://doi.org/10.1093/acprof:oso/9780199599844.003.0005>.

84. Nita A. Farahany & James E. Coleman, Jr., *Genetics, Neuroscience, and Criminal Responsibility*, in *The Impact of Behavioral Sciences on Criminal Law* 183 (Nita A. Farahany ed., 2009), <https://doi.org/10.1093/acprof:oso/9780195340525.003.0007>.

85. Oliver Wendell Holmes, Jr., *The Common Law* 66 (1881).

are undergoing the neurological evaluation.⁸⁶ Such neurological evaluation might even show that the defendant had a stroke or tumor in a particular part of the brain, which then could be used to argue in some way against the defendant's criminal responsibility.⁸⁷

Neuroimaging has been used more broadly in some criminal cases. For example, in the trial of John Hinckley Jr. for the attempted assassination of President Reagan, the defense used CAT scans of Hinckley's brain to support the argument, based largely on his bizarre behavior, that he suffered from schizophrenia. The scientific basis for that conclusion, offered early in the history of brain CAT scans, was questionable at the time and has become even weaker since, but Hinckley was found not guilty by reason of insanity. In November 2009, testimony about an fMRI scan was introduced in the penalty phase of a capital case as mitigating evidence that the defendant suffered from psychopathy. The defendant was sentenced to death, but after longer jury deliberations than defense counsel expected.⁸⁸ (This appears to have been the first time fMRI results were introduced in a criminal case.⁸⁹) To date, the most common types of neuroimaging that are introduced to support a defendant's claim about a brain abnormality or difference that at least in part explains their criminal conduct have been MRI or PET scans, rather than more advanced techniques like EEG, SPECT, or fMRI scanning.⁹⁰

Neuroscience evidence also may be relevant in wider arguments about criminal justice. Evidence about the development of adolescent brains has been referred to in appellate cases concerning the punishments appropriate for people who committed crimes while underage, including, as noted above, U.S. Supreme Court decisions. More broadly, some have urged that neuroscience will undercut much of the criminal justice system. The argument is that neuroscience

86. See, e.g., Nita A. Farahany, *Neuroscience and Behavioral Genetics in US Criminal Law: An Empirical Analysis*, 2 J.L. & Biosciences 485 (2015), <https://doi.org/10.1093/jlb/lsv059>.

87. In at least one fascinating case, a man who was convicted of sexual abuse of a child was found to have a large tumor pressing into his brain. When the tumor was removed, his criminal sexual impulses disappeared. When his impulses returned, so had his tumor. The tumor was removed a second time and, again, his impulses disappeared. Jeffrey M. Burns & Russell H. Swerdlow, *Right Orbitofrontal Tumor with Pedophilia Symptom and Constructional Apraxia Sign*, 60 Archives Neurology 437 (2003), <https://perma.cc/X2VU-VHWC>; *Doctors Say Pedophile Lost Urge After Tumor Removed*, USA Today, July 28, 2003, <https://perma.cc/NY89-2GQT>; see also Greely, *supra* note 83 (offering a longer discussion of this case).

88. The defendant in this Illinois case, Brian Dugan, confessed to the murder but sought to avoid the death penalty. See Virginia Hughes, *Science in Court: Head Case*, 464 Nature 340 (2010), <https://doi.org/10.1038/464340a> (providing an excellent discussion of this case).

89. It should be noted that other forms of neuroimaging, particularly PET and structural MRI scans, have been more widely used in criminal cases. Dr. Ruben Gur at the University of Pennsylvania estimates that he has used neuroimaging in testimony for criminal defendants about thirty times. *Id.*

90. Farahany, *supra* note 86.

ultimately will prove that no one—not even the sanest defendant—has free will and that this will fatally weaken the retributive aspect of criminal justice.⁹¹

The application of neuroscience evidence to individual claims of a lack of criminal responsibility has proven challenging.⁹² Such claims suffer from the time machine problem—the neuropsychological testing or imaging will almost always be from after, usually long after, the crime was committed and so cannot directly show a defendant’s brain state (and hence, by inference, his or her mental state) at or before the time of the crime. While this could change with more passively collected brain data if consumer neurotechnology becomes widespread, it remains to be seen how much we will be able to infer from that data, or how robust or reliable the data will be in an environment that is not well controlled, like a laboratory or clinical setting.

Careful neuroscience studies, either structural or functional, of the brains of criminals are rare. It seems highly unlikely that a “responsibility region” will ever be found, one that is universally activated in law-abiding people and that is deactivated in criminals (or vice versa). At most, the evidence is likely to show that people with particular brain structures or patterns of brain functioning commit crimes more frequently than people without such structures or patterns. Applying this group evidence to individual cases will be difficult, if not impossible. All of the problems of technical and statistical analysis of neuroimaging data, discussed above in the section titled “Issues in Interpreting Study Results,” apply. And it is possible that the to-be-scanned defendants will be able to implement countermeasures to “fool” the expert analyzing the scan.

The use of neuroscience to undermine criminal responsibility faces another problem: identifying a specific legal argument. It is not generally a defense to a criminal charge to assert that one has a predisposition to commit a crime, or even a very high statistical likelihood, as a result of social and demographic variables, of committing a crime. It is not clear whether neuroscience would, in any

91. See, e.g., Robert M. Sapolsky, *The Frontal Cortex and the Criminal Justice System*, in *Law and the Brain* 227 (Semir Zeki & Oliver Goodenough eds., 2006), <https://doi.org/10.1093/oso/9780198570103.003.0012>; Joshua Greene & Jonathan Cohen, *For the Law, Neuroscience Changes Nothing and Everything*, in *id.* at 207, <https://doi.org/10.1093/oso/9780198570103.003.0011>. These arguments have been forcefully made by Professor Stephen Morse. See, e.g., Stephen J. Morse, *Determinism and the Death of Folk Psychology: Two Challenges to Responsibility from Neuroscience*, 9 Minn. J.L. Sci. & Tech. 1 (2008), <https://perma.cc/985T-B4MR>; Stephen J. Morse, *The Non-Problem of Free Will in Forensic Psychiatry and Psychology*, 25 Behav. Sci. & L. 203 (2007), <https://doi.org/10.1002/bsl.744>; Stephen J. Morse, *Moral and Legal Responsibility and the New Neuroscience*, in *Neuroethics: Defining the Issues in Theory, Practice, and Policy* 33 (Judy Illes ed., 2d ed. 2006), <https://doi.org/10.1093/acprof:oso/9780198567219.003.0003>; Stephen J. Morse, *Brain Overclaim Syndrome and Criminal Responsibility: A Diagnostic Note*, 3 Ohio St. J. Crim. L. 397 (2006).

92. A good short discussion of these challenges can be found in A Judge’s Guide to Neuroscience: A Concise Introduction 37, *supra* note 5.

more than a very few cases,⁹³ provide evidence that was not equivalent to predisposition evidence. (And, of course, prudent defense counsel might think twice before presenting evidence to the jury that their client was strongly predisposed to commit crimes.)

We are at an early stage in our understanding of the brain and of the brain states related to the mental states involved in criminal responsibility. At this point, about all that can be said is that at least some criminal defense counsel, seeking to represent their clients zealously, will watch neuroscience carefully for arguments they could use to relieve their clients from criminal responsibility.

Lie Detection

In General

The use of neuroscience methods for lie detection probably has received more attention than any other issue raised in this reference guide.⁹⁴ This is due in part

93. See Greely, *supra* note 83 (arguing for a very narrow neuroscience-based defense).

94. For a technology whose results have yet to be admitted in court, the legal and ethical issues raised by fMRI-based lie detection have been discussed in an amazingly long list of scholarly publications from 2004 to the present. An undoubtedly incomplete list follows: Brown & Murphy, *supra* note 36; Anthony D. Wagner, *supra* note 23; Frederick Schauer, *Can Bad Science Be Good Evidence?: Neuroscience, Lie-Detection, Beyond*, 95 Cornell L. Rev. 1191 (2010); Frederick Schauer, *Neuroscience, Lie-Detection, and the Law: A Contrarian View*, 14 Trends Cognitive Scis. 101 (2010), <https://doi.org/10.1016/j.tics.2009.12.004>; Bizzi et al., *supra* note 23; Joëlle Anne Moreno, *The Future of Neuroimaged Lie Detection and the Law*, 42 Akron L. Rev. 717 (2009), <https://perma.cc/E58C-H7PQ>; Julie Seaman, *Black Boxes: fMRI Lie Detection and the Role of the Jury*, 42 Akron L. Rev. 931 (2009), <https://perma.cc/JHW9-RGP2>; Jane Campbell Moriarty, *Visions of Deception: Neuroimages and the Search for Truth*, 42 Akron L. Rev. 739 (2009), <https://perma.cc/SY9C-N24G>; Benjamin Holley, *It's All in Your Head: Neurotechnological Lie Detection and the Fourth and Fifth Amendments*, 28 Devs. Mental Health L. 1 (2009), <https://perma.cc/2GWS-WWPZ>; Brian Reese, Comment, *Using fMRI as a Lie Detector—Are We Lying to Ourselves?*, 19 Alb. L.J. Sci. & Tech. 205 (2009), <https://perma.cc/XJ2D-4QQF>; Cooper Ellenberg, *Lie Detection: A Changing of the Guard in the Quest for Truth in Court?*, 33 L. & Psych. Rev. 139 (2009), <https://perma.cc/R58K-58ZX>; Julie A. Seaman, *Black Boxes*, 58 Emory L.J. 427 (2008), <https://perma.cc/PAL6-B4BY>; Greely & Illes, *supra* note 44; Mark Pettit, *fMRI and BF Meet FRE: Brain Imaging and the Federal Rules of Evidence*, 33 Am. J.L. & Med. 319 (2007), <https://doi.org/10.1177/009885880703300208>; Jonathan H. Marks, *Interrogational Neuroimaging in Counterterrorism: A “No-Brainer” or a Human Rights Hazard?*, 33 Am. J.L. & Med. 483 (2007), <https://perma.cc/R2TJ-WNCS>; Leo Kittay, *Admissibility of fMRI Lie Detection: The Cultural Bias Against “Mind Reading” Devices*, 72 Brook. L. Rev. 1351, 1355 (2007), <https://perma.cc/33D7-86TA>; Jeffrey Bellin, *The Significance (If Any) for the Federal Criminal Justice System of Advances in Lie Detector Technology*, 80 Temp. L. Rev. 711 (2007), <https://perma.cc/XNV3-RE36>; Henry T. Greely, *The Social Consequences of Advances in Neuroscience: Legal Problems, Legal Perspectives*, in *Neuroethics: Defining the Issues in Theory, Practice, and Policy* 245, *supra* note 91, <https://doi.org/10.1093/acprof:oso/9780198567219.003.0017>; Charles N.W. Keckler, *Cross-Examining the Brain: A Legal Analysis of Neural Imaging for Credibility Impeachment*, 57 Hastings L.J. 509 (2006), <https://perma.cc/QVY5-7ZSR>; Archie Alexander, *Functional Magnetic Resonance Imaging Lie Detection: Is a “Brainstorm” Heading*

to the cultural interest in lie detection, dating back in its technological phase nearly a hundred years to the invention of the polygraph.⁹⁵ But it is also due to the fact that for several years two commercial firms offered fMRI-based lie-detection services for sale in the United States: Cephos and No Lie MRI.⁹⁶

Beyond the scientific validity of these techniques lies a host of legal questions. How accurate is accurate enough for admissibility in court or for other legal system uses? What are the implications of admissible and accurate lie detection for the Fourth, Fifth, Sixth, and Seventh Amendments? Would jurors be allowed to consider the failure, or refusal, of a party to take a lie-detector test? Would lie detection be available in discovery? Would each side get to do its own tests—and who would pay?

Accurate lie detection could make the justice system much more accurate. Incorrect convictions might become rare; so might incorrect acquittals. Accurate lie detection also could make the legal system much more efficient. It seems likely that far fewer cases would go to trial if the witnesses could expect to have their veracity accurately determined.

Inaccurate lie detection, on the other hand, holds the potential of ruining the innocent and immunizing the guilty. It is at least daunting to remember some of the failures of the polygraph, such as the case of Aldrich Ames, a Soviet (and then Russian) mole in the Central Intelligence Agency, who passed two Agency polygraph tests while serving as a paid spy.⁹⁷ The courts already have begun to decide whether and how to use these new methods of lie detection in

Toward the “Gatekeeper”?, 7 Hous. J. Health L. & Pol’y 1 (2007), <https://perma.cc/H4UR-HBR2>; Paul Root Wolpe, Kenneth R. Foster & David D. Langleben, *Emerging Neurotechnologies for Lie-Detection: Promises and Perils*, 5 Am. J. Bioethics 39, 42 (2005), <https://doi.org/10.1080/1526516059023367>; Henry T. Greely, *Premarket Approval Regulation for Lie Detection: An Idea Whose Time May Be Coming*, 5 Am. J. Bioethics, 50–52 (2006), <https://doi.org/10.1080/15265160590960988>; Sean Kevin Thompson, Note, *The Legality of the Use of Psychiatric Neuroimaging in Intelligence Interrogation*, 90 Cornell L. Rev. 1601 (2005), <https://perma.cc/DC5N-PFBK>; Henry T. Greely, *Prediction, Litigation, Privacy, and Property: Some Possible Legal and Social Implications of Advances in Neuroscience*, in *Neuroscience and the Law* 114–56, *supra* note 5, <https://perma.cc/9JXM-PVMV>; Judy Illes, *A Fish Story? Brain Maps, Lie Detection, and Personhood*, 6 Cerebrum 73 (2004), <https://perma.cc/8ZY4-V7X6>.

95. An interesting history of the polygraph can be found in Ken Alder, *The Lie Detectors: The History of an American Obsession* (2007). Perhaps the best overall discussion of the polygraph, including some discussion of its history, is found in the National Research Council report, commissioned in the wake of the Wen Ho Lee case, on the use of the technology for screening. Nat’l Rsch. Council, *supra* note 25.

96. Cephos stopped providing fMRI-based lie-detection services sometime after 2012; it seems to have some continuing existence with respect to DNA forensics. *See* Cephos, <https://perma.cc/F3JY-9D73>. No Lie MRI appears to have changed its name to Truthful Brain Corporation and to have created a nonprofit venture called Medicine for Law. It is not clear whether either of those is still active, although the second does have a website: <https://perma.cc/V78N-EB3F>.

97. *See* Senate Select Committee on Intelligence, *An Assessment of the Aldrich H. Ames Espionage Case and Its Implications for U.S. Intelligence* (1994), <https://perma.cc/26TT-55SK>.

the judicial process; the rest of society also will soon be forced to decide on their uses and limits.

But, of course, lie detection might have applications to litigation without ever being introduced in trials. As is the case today with the polygraph, the fact that it is not generally admissible in court might not stop the police or the prosecutors from using it to investigate alleged crimes. Similarly, defense counsel might well use it to attempt to persuade the authorities that their clients should not be charged or should be charged with lesser offenses. One could imagine the same kinds of pretrial uses of lie detection in civil cases, as the parties seek to affect each other's perceptions of the merits of the case.

Such lie-detection efforts could also affect society, and the law, outside of litigation. One could imagine prophylactic lie detection at the beginning of contractual relations, seeking to determine whether the other side honestly had the present intention of complying with the contract's terms. One can also imagine schools using lie detection as part of investigations of student misconduct, or parents seeking to use lie detection on their children. The law more broadly may have to decide whether and how private actors can use lie detection, determining whether, for example, to extend to other contexts—or to weaken or repeal—the Employee Polygraph Protection Act.⁹⁸

Efforts to use fMRI-based lie detection in the courts have faded in recent years, but another method of “recognition detection” may soon reach the courts. This section will discuss each approach, both to help judges if such evidence reaches their courts but also, at least for fMRI-based lie detection, as a case study of the legal system's response to such evidence.

fMRI-Based Lie Detection

Currently, as far as we know, evidence from fMRI-based lie detection has not been admitted into evidence in any court, but it was offered—and rejected—in at least three cases, *United States v. Semrau*,⁹⁹ *Wilson v. Corestaff Services L.P.*,¹⁰⁰ and a widely publicized murder trial, *Maryland v. Smith*,¹⁰¹ in the early 2010s.¹⁰²

98. See *supra* discussion accompanying note 54.

99. 693 F.3d 510 (6th Cir. 2012), *aff'g*, No. 1:07CR10074-01-MI, 2011 WL 1114441 (W.D. Tenn. Mar. 24, 2011).

100. 900 N.Y.S.2d 639 (N.Y. Sup. Ct. 2010).

101. Michael Laris, *Debate on Brain Scans as Lie Detectors Highlighted in Maryland Murder Trial*, Wash. Post, Aug. 26, 2012, <https://perma.cc/EG4Z-FZ84>.

102. In early 2009, a motion to admit fMRI-based lie-detection evidence, provided by No Lie MRI, was made, and then withdrawn, in a child custody case in San Diego. The case is discussed in a prematurely titled article by Alexis Madrigal, *fMRI Lie Detection to Get First Day in Court*, WIRED Sci., Mar. 16, 2009, <https://perma.cc/29SF-VKFH>. A somewhat similar method of using EEG to look for signs of “recognition” in the brain was admitted into one state court hearing for

Since then, efforts to introduce evidence from this technology have largely disappeared—we find no reported cases about such attempted uses since 2012—but the story of its rise and fall provides a good example of a path for new neuroscience evidence. And the underlying science still exists and could well end up in court again. This section will begin by analyzing the issues raised for courts by this technology and then will discuss these three cases, before ending with a quick look at possible uses for this kind of technology outside the courtroom.

Published research on fMRI and detecting deception dates back to about 2001.¹⁰³ As noted above, to date between twenty and thirty peer-reviewed articles from about fifteen laboratories have appeared claiming to find statistically significant correlations between patterns of brain activation and deception. Only a handful of the published studies have looked at the accuracy of determining deception in individual participants as opposed to group averages. Those studies generally claim accuracy rates of between about 75% and 90%. No Lie MRI has licensed the methods used by one laboratory, that of Dr. Daniel Langleben at the University of Pennsylvania; Cephos has licensed the method used by another laboratory, that of Dr. Frank A. Kozel, first at the Medical University of South Carolina and then at the University of Texas Southwestern Medical Center. (The method used by a British researcher, Dr. Sean Spence, has been used on a British reality television show.)

All of these studies rely on research participants, typically but not always college students, who are recruited for a study of deception. They are instructed to answer some questions truthfully in the scanner and to answer other questions inaccurately.¹⁰⁴ In the Langleben studies, for example, undergraduates were shown images of playing cards while in the scanner and asked to indicate whether they saw a particular card. They were instructed to answer truthfully except when they saw one particular card. Some of Kozel's studies used a different experimental paradigm, in which the participants were put in a room and told to take either a watch or a ring. When asked in the scanner separately whether they

postconviction relief at the trial court level in Iowa in 2001, and both it and another EEG-based method have been used in India. As far as we know, evidence from the use of EEG for lie detection has not been admitted in any other U.S. cases. *See supra* note 45.

103. The most recent reviews of the scientific literature on this subject are Wagner, *supra* note 23, and Shawn E. Christ et al., *The Contributions of Prefrontal Cortex and Executive Control to Deception: Evidence from Activation Likelihood Estimate Meta-Analyses*, 19 *Cerebral Cortex* 1557 (2009), <https://doi.org/10.1093/cercor/bhn189>. *See also* Greely & Illes, *supra* note 44 (for discussion of the articles through early 2007). The following discussion is based largely on those sources.

104. At least one fMRI study has attempted to investigate self-motivated lies, told by participants who were *not* instructed to lie, but who chose to lie for personal gain. Joshua D. Greene & Joseph M. Paxton, *Patterns of Neural Activity Associated with Honest and Dishonest Moral Decisions*, 106 *PNAS* 12506 (2009), <https://doi.org/10.1073/pnas.0900152106>. The experiment was designed to make it easy for participants to realize they would be given more money if they lied about how many times they correctly predicted a coin flip. Investigators could not, however, determine if a participant lied in any particular trial.

had taken the watch and then whether they had taken the ring, they were to reply “no” in both cases—truthfully once and falsely the other time. When analyzed in various ways, the fMRI results showed statistically different patterns of brain activation (small changes in BOLD response) when the participants were lying and when they were telling the truth.

In general, these studies are not guided by a consistent hypothesis about which brain regions should be activated or deactivated during truth or deception. The results are empirical; they see particular patterns that differ between the truth state and the lie state. Some have argued that the patterns show greater mental effort when deception is involved; others have argued that they show more impulse control when lying.

Are fMRI-based lie-detection methods accurate? As a class of experiments, these studies are subject to all the general problems discussed above in the section “Issues in Interpreting Study Results” regarding fMRI scans that might lead to neuroscience evidence. So far, there are only a few studies involving a limited number of participants. (The method used by No Lie MRI seems ultimately to have been based on the responses of four right-handed, healthy, male University of Pennsylvania undergraduates.¹⁰⁵) There have been, to date, no independent replications of any group’s findings.

The experience of the research participants in these fMRI studies of deception seems to be different from “lying” as the court system would perceive it. The participants knew they were involved in research, they were following orders to lie, and they knew that the most harm that could come to them from being detected in a lie might be lesser payment for taking part in the experiment. This seems hard to compare to a defendant lying about participating in a murder.¹⁰⁶ More fundamentally, it is not clear how one could conduct ethical but realistic experiments with lie detection. Research participants cannot credibly be threatened with jail if they do not convince the researcher of the truth of their lies.

Only a handful of researchers have published studies showing reported accuracy rates with individual participants and only with a small number of participants.¹⁰⁷ Some of the studies used complex and somewhat controversial

105. Langleben et al., *supra* note 23, at 267.

106. Michael Pardo has forcefully made this point by dissecting, philosophically, the meaning of the term *lie*. In 2018, in one of the latest scholarly discussions of fMRI lie detection, he argues that the studies are not detecting “lying” because they examine participants who are doing what they have been instructed to do. Michael S. Pardo, *Lying, Deception, and fMRI: A Critical Update*, in *Neurolaw and Responsibility for Action* 143 (Bebhinn Donnelly-Lazarov ed., 2018), <https://perma.cc/FLD9-PA5U>. Following earlier work from 2013, see Michael S. Pardo & Dennis Patterson, *Minds, Brains, and Law: The Conceptual Foundations of Law and Neuroscience* 105–06 (2013), where he discusses five subsequent studies, all of which he finds unhelpful.

107. See discussion in Wagner, *supra* note 23, at 29–35. Wagner analyzes eleven peer-reviewed, published papers. Seven come from Kozel’s laboratory; three come from Langleben’s. The only exception is a paper from John Cacioppo’s group, which concludes “[A]lthough fMRI may permit

statistical techniques. And although participants in at least one experiment were invited to try to use countermeasures against being detected, no specific countermeasures were tested.

United States v. Semrau is the most significant case involving fMRI lie detection. On June 1, 2010, U.S. Magistrate Judge Tu M. Pham of the Western District of Tennessee issued an amended thirty-nine-page report and recommendation on the prosecution's motion to exclude evidence from an fMRI-based lie-detection report by Cephos.¹⁰⁸ The report came after a hearing on May 13–14 featuring testimony from Steve Laken, CEO of Cephos, for admission, and from two experts arguing against admission. The district judge adopted the magistrate's report in its entirety during the trial and, in September 2012, the Sixth Circuit, in a lengthy opinion that relied heavily on the magistrate judge's report and recommendations, affirmed that the district court had not abused its discretion in excluding the evidence under either Rule 702 or Rule 403.¹⁰⁹

The defendant in this case, Dr. Semrau, a health professional accused of defrauding Medicare, offered as evidence a report from Cephos stating that he was being truthful when he answered a set of questions about his actions and knowledge concerning the alleged crimes.

Judge Pham first analyzed the motion under Rule 702, using the *Daubert* criteria. He concluded that the technique was testable and had been the subject of peer-reviewed publications. On the other hand, he concluded that the error rates for its use in realistic situations were unknown. Furthermore, he found there were no standards for its appropriate use. To the extent that the publications relied on by Cephos to establish its reliability constituted such standards, those standards had not actually been followed in the tests of the defendant. Cephos actually scanned Dr. Semrau two times on one day, asking questions about one aspect of the criminal charges during the first scan and then about another aspect in the second scan. The company's subsequent analysis of those scans indicated that the defendant had been truthful in the first scan but deceptive in the second scan. Cephos then scanned him a third time, several days later, on the second subject but with revised questions, and concluded that he was telling the truth that time. Nothing in the publications relied upon by Cephos indicated that the third scan

investigation of the neural correlates of lying, at the moment it does not appear to provide a very accurate marker of lying that can be generalized across individuals or even perhaps across types of lies by the same individuals." George T. Monteleone et al., *Detection of Deception Using fMRI: Better Than Chance, But Well Below Perfection*, 4 Soc. Neuroscience 528 (2009), <https://doi.org/10.1080/17470910801903530>. However, that study only looked at one brain region at a time, and it did not test combinations or patterns, which might have improved the predictive power.

108. *United States v. Semrau*, No. 07–10074, 2010 WL 6845092 (W.D. Tenn. June 1, 2010). The district court judge assigned to the case had a scheduling conflict on the date of the hearing on the prosecution's motion, so the hearing was held before a magistrate judge from that district.

109. 693 F.3d 510, 523–24 (6th Cir. 2012).

was appropriate. Finally, Judge Pham found that the method was not generally accepted in the relevant scientific community as sufficiently reliable for use in court, citing several publications, including some written by the authors whose methods Cephos used.

The magistrate judge then examined the motion under Rule 403 and found that the potential prejudicial effect of the evidence outweighed its probative value. He noted that the test had been conducted without the government's knowledge or participation, in a context where the defendant risked nothing by taking the test—a negative result would never be disclosed. He noted the jury's central role in determining credibility and considered the likelihood that the lie-detection evidence would be a lengthy and complicated distraction from the jury's central mission. Finally, he noted that the probative value of the evidence was greatly reduced because the report only gave a result concerning the defendant's general truthfulness when responding to more than ten questions about the events but did not even purport to say whether the defendant was telling the truth about any particular question.¹¹⁰

The Sixth Circuit panel affirmed unanimously, holding that the district court did not abuse its discretion in excluding the fMRI evidence pursuant to Rule 403 in light of (1) the questions surrounding the reliability of fMRI lie-detection tests in general and as performed on Dr. Semrau, (2) the failure to give the prosecution an opportunity to participate in the testing, and (3) the test result's inability to corroborate Dr. Semrau's answers as to the particular offenses for which he was charged.¹¹¹

In the same month as the magistrate judge's report in *Semrau*, a state trial court judge in Brooklyn excluded another Cephos lie-detection report in a civil case, *Wilson v. Corestaff Services L.P.*¹¹² This case involved a claim under state law by a former employee that she had been subject to retaliation for reporting sexual harassment. The plaintiff offered evidence from a Cephos report finding that

110. Interestingly, the Sixth Circuit panel stated, in a footnote, that “the prospect of introducing fMRI lie detection results into criminal trials is undoubtedly intriguing and, perhaps, a little scary.” *Id.* at 524 n.12. See Daniel S. Goldberg, *Against Reductionism in Law & Neuroscience*, 11 Hous. J. Health L. & Pol'y 321, 324 & n.6 (2011), <https://perma.cc/K4E4-6J4Q> (reviewing literature that “challenges the very idea that fMRI or other novel neuroimaging techniques either can or should be used as evidence in criminal proceedings”).

There may well come a time when the capabilities, reliability, and acceptance of fMRI lie detection—or even a technology not yet envisioned—advances to the point that a trial judge will conclude, as did Dr. Laken in this case: “I would subject myself to this over a jury any day.” Though we are not at that point today, we recognize that as science moves forward the balancing of Rule 403 may well lean toward finding that the probative value for some advancing technology is sufficient.

Semrau, 693 F.3d at 524 n.12.

111. *Id.* at 524.

112. 900 N.Y.S.2d 639 (N.Y. Sup. Ct. 2010).

her main witness was truthful when he described how the defendant's management said it would retaliate against the plaintiff.

That case did not involve an evidentiary hearing or, indeed, any expert testimony. The judge decided the lie-detection evidence was not appropriate under New York's version of the *Frye* test, noting that, in New York, "courts have advised that the threshold question under *Frye* in passing on the admissibility of expert's testimony is whether the testimony is 'within the ken of the typical juror.'" ¹¹³ Because credibility is a matter for the jury, the judge concluded that this kind of evidence was categorically excluded under New York's version of *Frye*. He also noted that "even a cursory review of the scientific literature demonstrates that the plaintiff is unable to establish that the use of the fMRI test to determine truthfulness or deceit is accepted as reliable in the relevant scientific community." ¹¹⁴

Finally, also in 2012, a Maryland trial court excluded a No Lie MRI fMRI-based lie-detection report in a well-publicized murder case. The judge held three days of pretrial hearings on the evidence before concluding there was no consensus among the experts: "These are brilliant people, and they don't agree." ¹¹⁵ No published opinion followed. ¹¹⁶

After 2012, fMRI-based lie detection continued to be discussed in the press and the academic literature for several years, ¹¹⁷ but without any other known efforts to introduce it. (*Semrau*, for example, has yet to be cited in any opinions involving lie detection.) Nonetheless, the science is still there and further attempts may be made.

The current fMRI-based methods of lie detection do provide one kind of protection for possible participants—they are obvious. No one is going to be put into an MRI for an hour and asked to respond, repeatedly, to questions without realizing something important is going on. Should researchers develop less obtrusive or obvious methods of neuroscience-based lie detection, we will have to deal with the possibilities of involuntary and, indeed, surreptitious lie detection.

113. *Id.* at 642 (citing *People v. Cronin*, 60 N.Y.2d 430, 433 (1983)).

114. *See id.*

115. Michael Laris, *Debate on Brain Scans as Lie Detectors Highlighted in Maryland Murder Trial*, Wash. Post, Aug. 26, 2012, <https://perma.cc/X5G9-GVPT>.

116. A subsequent appeal was published, *Smith v. State*, 98 A.3d 444 (Md. Ct. Spec. App. 2014), but the lie-detection evidence had not been a ground for that appeal and was not mentioned.

117. Notable additions to the scholarly literature, in addition to the Pardo works cited *supra* note 106, include Daniel D. Langleben & Jane Campbell Moriarty, *Using Brain Imaging for Lie Detection: Where Science, Law, and Policy Collide*, 19 Psych., Pub. Pol'y & L. 222 (2013), <https://doi.org/10.1037/a0028841>; Martha J. Farah et al., *Functional MRI-based Lie Detection: Scientific and Societal Challenges*, 15 Nature Revs. Neuroscience 123 (2014), <https://doi.org/10.1038/nrn3665>; Jane Campbell Moriarty, *Neuroimaging Evidence in US Courts*, in *Law and Mind: A Survey of Law and the Cognitive Sciences* 370 (Bartosz Brożek, Jaap Hage & Nicole A. Vincent eds., 2021), <https://perma.cc/YR36-LGYU>.

Issues Involved in the Use of EEG-Based Brain “Recognition”

A different approach uses the unique neural signature in EEGs of suspects’ brains associated with the recognition, salience, or congruence of facts or events to interrogate the brain.¹¹⁸ While criminal suspects wear an EEG headset, they are shown images, words, or sounds during which their brainwave activity is recorded. Images from a crime scene such as a murder weapon, objects from crime scenes, the victim, or the victim’s clothing might all be triggering images. An analyst then interprets the brainwave activity to see if a characteristic EEG signal registers from the brain in response to any of the images.

The P300 wave was the first to be used for this purpose. It is an event-related potential (ERP) measurement of brain activity, where the amplitude of that waveform differs depending on whether a participant recognizes a prior fact.¹¹⁹ Whereas a traditional polygraph asks a person a series of yes or no questions and then looks at their physiological responses to infer if they are telling the truth or a lie, this approach is designed to reveal “recognition” of salient facts of a crime.¹²⁰

Although this EEG-based technique was initially heralded as an extraordinary scientific breakthrough, the broader scientific community raised serious doubts about the veracity of the approach.¹²¹ More recently, however, some scientists have started to independently test the technique and are encouraged that it could be applied more reliably.¹²² In particular, Dr. Peter Rosenfeld rigorously tested this approach, including looking for countermeasures—and countermeasures to countermeasures.¹²³

Another scientifically promising approach uses a different ERP brain response called the N400 signal, which seems to respond to congruence rather than salience. As an author of this reference guide explained:

You could show a suspect a series of faces that includes their suspected co-conspirators. Their N400 brain signals would be more negative for “incongruent” faces that didn’t belong than for “congruent” faces that did. Similarly, you could pair words together like “body” and “lake” versus “body” and “base-ment” to try to find out where a murder victim’s corpse is hidden.¹²⁴

118. See Farahany, *supra* note 2.

119. Tim Stelloh, *Larry Farwell Claims His Lie Detector System Can Read Your Mind. Is He a Scam Artist, or a Genius?*, Medium, Jan. 6, 2021, <https://perma.cc/N9S7-YU9K>.

120. Farahany, *supra* note 43.

121. Stelloh, *supra* note 119.

122. *Id.*

123. See J. Peter Rosenfeld, *P300 in Detecting Concealed Information and Deception: A Review*, 57 *Psychophysiology* e13362 (2020), <https://doi.org/10.1111/psyp.13362>.

124. Farahany, *supra* note 2, at 82. See also K. V. Dijkstra, J. D. R. Farquhar & P. W. M. Desain, *The N400 for Brain Computer Interfacing: Complexities and Opportunities*, 17 *J. Neural Eng’g* 022001 (2020), <https://doi.org/10.1088/1741-2552/ab702e>.

Both the P300 and the N400 signals are detected with EEG, a cheap and easily used technology. Some of the earliest attempts in the United States to use the P300 technology, oddly referred to as “brain fingerprinting,” came from criminal defendants themselves, asking for the results of the test to be used to validate their claims of innocence.¹²⁵ More commonly, it is the police and not criminal suspects who have sought out the P300 technology.¹²⁶ But the practice has not become widespread in the United States, likely because criminal defendants will only rarely voluntarily submit to the test and cannot now be forced to do so. But it is finding some success overseas. P300 has been used in India since 2003. Singapore’s police services purchased the technology in 2013; and one source says police in Florida signed a contract in 2014 to use the approach.¹²⁷ A firm now named Brainwave Science has been marketing this approach, calling it “iCognitive,” with some success around the world. Australian counterterrorism authorities are currently looking into the potential use of brain fingerprinting to determine whether individuals returning from war zones have been illegally involved in conflict when they claim to have been participating in humanitarian work.¹²⁸

Detection of Pain

Detection of pain is likely to be another area for use of neuroscience evidence, although this will be mainly in civil cases or administrative appeals (often from denial of government disability payments). No matter where an injury occurs and no matter where it seems to hurt, pain is felt in the brain.¹²⁹ Without sensory nerves leading to the brain from a body region, there is usually no experience of pain. Without the brain machinery and functioning to process the signal, no pain is perceived.

Pain turns out to be complicated—even the common pain that is experienced from an acute injury to, say, an arm. Neurons near the site of the injury called nociceptors transmit the pain signal to the spinal cord, which relays it to the brain. But other neurons near the site of the injury will, over time, adapt to affect the pain signal. Cells in the spinal cord can also modulate the pain signal that is sent to the brain, making it stronger or weaker. The brain, in turn,

125. *State v. Harrington*, 284 N.W.2d 244 (Iowa 1979); e.g., *Johnson v. State*, 730 N.W.2d 209 (Table) (Iowa Ct. App. 2007); *People v. Dorris*, 2013 IL App (4th) 120699–U (Ill. App. Ct. 2013).

126. Jayanth Murali, *Cool Tool for Police Investigation?*, *Deccan Chron.*, Sept. 2, 2018, <https://perma.cc/J2YY-X6SB>.

127. David Cox, *Can Your Brain Reveal You Are a Liar?*, *BBC*, Jan. 25, 2016, <https://perma.cc/NWR7-PAEV>.

128. *Id.*

129. See Brain Facts, *supra* note 6, at 19–21, 49–50, which includes a useful brief description of the neuroscience of pain.

sends signals down to the spinal cord that cause, or at least affect, these modulations. And the actual sensation of pain—the “ouch”—takes place in the brain.

The immediate and localized sensation is processed in the somatosensory cortex, the brain region that takes sensory inputs from different body parts (with each body part getting its own portion of the somatosensory cortex) and processes them into a perceived sensation. The added knowledge that the sensation is painful seems to require the participation of other regions of the brain. Using fMRI and other techniques, some researchers have identified what they call the “pain matrix” in the brain, regions that are activated when experimental participants, in scanners, are exposed to painful stimuli. The brain regions identified as part of the so-called pain matrix vary from researcher to researcher, but generally include the thalamus, the insula, parts of the anterior cingulate cortex, and parts of the cerebellum.¹³⁰

Researchers have run experiments with participants in the scanner receiving painful or nonpainful stimuli and have attempted to find activation patterns that appear when pain is perceived and that do not appear when pain is absent. (The participants usually are given nonharmful painful stimuli such as having their skin touched with a hot metal rod or coated with a pepper-derived substance that causes a burning sensation.) Some have reported substantial success, detecting pain in more than 80% of the cases.¹³¹ Other studies have found a positive correlation between the degree of activation in the pain matrix and the degree of subjective pain, both as reported by the participants and as possibly indicated by the heat of the rod or the amount of the painful substance—the higher the temperature or the concentration of the painful substance, the greater the average activation in the pain matrix.¹³²

Other neuroscience studies of individual pain look not at brain function during painful episodes but at brain structure. Some researchers, for example, claim that different regions of the brain have different average size and neuron densities in patients who have had long-term chronic pain than in those who have not had such pain.¹³³

130. A good recent review of this work can be found in Maite M. van der Miesen, Martin A. Lindquist & Tor D. Wager, *Neuroimaging-based Biomarkers for Pain: State of the Field and Current Directions*, 4 PAIN Reps. e751 (2019), <https://doi.org/10.1097/PR.0000000000000751>.

131. See, e.g., Irene Tracey, *Imaging Pain*, 101 Brit. J. Anaesthesia 32 (2008), <https://doi.org/10.1093/bja/aen102>.

132. See, e.g., Robert C. Coghill, John G. McHaffie & Yi-Fen Yen, *Neural Correlates of Interindividual Differences in the Subjective Experience of Pain*, 100 PNAS 8538 (2003), <https://doi.org/10.1073/pnas.1430684100>.

133. See, e.g., A. Vania Apkarian et al., *Chronic Back Pain Is Associated with Decreased Prefrontal and Thalamic Gray Matter Density*, 24 J. Neuroscience 10410 (2004), <https://doi.org/10.1523/JNEUROSCI.2541-04.2004>; see also Arne May, *Chronic Pain May Change the Structure of the Brain*, 137 Pain 7 (2008), <https://doi.org/10.1016/j.pain.2008.02.034>; Karen D. Davis, *Recent Advances and Future Prospects in Neuroimaging of Acute and Chronic Pain*, 1 Future Neurology 203 (2006), <https://doi.org/10.2217/14796708.1.2.203>.

Pain is clearly complicated. Placebos, distractions, or great need can sometimes cause people not to sense, or perhaps not to notice, pain that could otherwise be overwhelming. Similarly, some people can become hypersensitive to pain, reporting severe pain when the stimulus normally would be benign. Amputees with phantom pain—the feeling of pain in a limb that has been gone for years—have been scanned while reporting this phantom pain. They show activation in the pain matrix. In some fMRI studies, people who have been hypnotized to feel pain, even when there is no painful stimulus, show activation in the pain matrix.¹³⁴ And in one fMRI study, participants who reported feeling emotional distress, as a result of apparently being excluded from a “game” being played among research participants, also showed, on average, statistically significant activation of the pain matrix.¹³⁵

Pain also plays an enormous role in the legal system.¹³⁶ The existence and extent of pain is a matter for trial in hundreds of thousands of injury cases each year. Perhaps more importantly, pain figures into uncounted workers’ compensation claims and Social Security disability claims. Pain is often difficult to prove, and the uncertainty of a jury’s response to claimed pain probably keeps much litigation alive. We know that the tests for pain currently presented to jurors, judges, and other legal decision-makers are not perfect. Anecdotes and the assessments of pain experts both are convincing that some nontrivial percentage of successful claimants are malingering and only pretending to feel pain; a much greater percentage may be exaggerating their pain.

A good test for whether a person is feeling pain, and, even better, a “scientific” way to measure the amount of that pain—at least compared to other pains felt by that individual, if not to pain as perceived by third parties—could help resolve a huge number of claims each year. If such pain detection were reliable, it would make justice both more accurate and more certain, leading to faster and cheaper resolution of many claims involving pain. The legal system, as well as the honest plaintiffs and defendants within it, would benefit.

134. Stuart W. Derbyshire et al., *Cerebral Activation During Hypnotically Induced and Imagined Pain*, 23 *NeuroImage* 392 (2004), <https://doi.org/10.1016/j.neuroimage.2004.04.033>.

135. Naomi I. Eisenberger et al., *Does Rejection Hurt? An fMRI Study of Social Exclusion*, 302 *Sci.* 290 (2003), <https://doi.org/10.1126/science.1089134>.

136. For analyses of the legal and ethical implications of using neuroimaging to detect pain, see Amanda C. Pustilnik, *Status and Surveillance in the Use of Brain-Based Pain Imaging in the Law*, in 1 *Developments in Neuroethics & Bioethics: Pain Neuroethics and Bioethics* 59, 61 (Daniel Z. Buchman & Karen D. Davis eds., 2018), <https://doi.org/10.1016/bs.dnb.2018.08.004>; Davis et al., *supra* note 72; Henry T. Greely, *Neuroscience, Mindreading, and the Courts: The Example of Pain*, 18 *J. Health Care L. & Pol’y* 171 (2015), <https://perma.cc/6KXM-PQQB>. The earliest, but still useful, discussion is found in Adam J. Kolber, *Pain Detection and the Privacy of Subjective Experience*, 33 *Am. J.L. & Med.* 433 (2007), <https://doi.org/10.1177/009885880703300212>. Kolber expands on that discussion in interesting ways in Adam J. Kolber, *The Experiential Future of the Law*, 60 *Emory L.J.* 585, 596–601 (2010), <https://perma.cc/P9QQ-AYWE>.

A greater understanding of pain also might lead to broader changes in the legal system. For example, emotional distress often is treated less favorably than direct physical pain. If neuroscience were to show that, in the brain, emotional distress seemed to be the same as physical pain, the law might change. Perhaps that would be more likely if neuroscience could provide assurance that sincere emotional pain could be detected and faked emotional distress would not be rewarded, the law again might change. Others have argued that even our system of criminal punishment might change if we could measure, more accurately, how much pain different punishments caused defendants, allowing judges to let the punishment fit the criminal, if not the crime.¹³⁷ A “pain detector” might even change the practice of medicine in legally relevant ways, by giving physicians a more certain way to check whether their patients are seeking controlled substances to relieve their own pain or are seeking them to abuse or to sell for someone else to abuse.

In at least one case, a researcher who studies the neuroscience of pain was retained as an expert witness to testify regarding whether neuroimaging could provide evidence that a claimant was, in fact, feeling pain. The case settled before the hearing.¹³⁸ In another case, a prominent neuroscientist was approached about being a witness against the admissibility of fMRI-based evidence of pain, but, before she had decided whether to take part, the party seeking to introduce the evidence changed its mind. This issue has not, as of the time of this writing, reached the courts yet, but lawyers clearly are thinking about these uses of neuroscience. (And note that in some administrative contexts, the evidentiary rules will not apply in their full rigor, possibly making the admission of such evidence more likely.)

Do either functional or structural methods of detecting pain work and, if so, how well? We do not know. These studies share many of the problems outlined above in the section titled “Issues in Interpreting Study Results.” The studies are few in number, with few participants (and usually sets of participants that are not very diverse). The experiments—usually involving giving college students a painful stimulus—are different from the experience of, for example, older people who claim to have lower back pain. Independent replication is rare, if it exists at all. The experiments almost always report that, on average, the group shows a statistically significant pattern of activation that differs depending on whether they are receiving the painful stimulus, but the group average does not in itself tell us about the sensitivity or specificity of such a test when applied to individuals. And the statistical and technical issues are daunting.

137. Adam J. Kolber, *How to Improve Empirical Desert*, 75 Brook. L. Rev. 433 (2009), <https://perma.cc/X56D-V8UR>.

138. Greg Miller, *Brain Scans of Pain Raise Questions for the Law*, 323 Sci. 195 (2009), <https://doi.org/10.1126/science.323.5911.195>.

In the area of pain, the issue of countermeasures may be the most interesting, particularly in light of the experiments conducted with hypnotized participants. Does remembered pain look the same in an fMRI scan as currently experienced pain? Does the detailed memory of kidney stone pain look any different from the present sensation of lower back pain? Can individuals effectively convince themselves that they are feeling pain and so appear to the scanner to be experiencing pain? The answer to these questions is clear: We do not yet know.

Pain detection also would raise legal questions. Could a plaintiff be forced to undergo a “pain scan”? If a plaintiff offered a pain scan in evidence, could the defendant compel the plaintiff to undergo such a scan with the defendant’s machine and expert? Would it matter if the scan were itself painful or even dangerous? Who would pay for these scans and for the experts to interpret them?

Detecting pain would be a form of neuroscience evidence with straightforward and far-reaching applications to the legal system. Whether it can be done, and, if so, how accurately, remain to be seen. So does the legal system’s reaction to this possibility.

Addiction

In addition to these three areas, we want to note another topic of great legal interest where neuroscience evidence may eventually play a major role: addiction. We are not giving it substantial discussion, because the research seems further from the courtroom at this point than with the other topics, but the extent of criminal cases involving addiction, in both federal and state courts, suggests its possible value.

In 1997, Alan Leshner, then director of the National Institute on Drug Abuse, published an influential article in *Science* titled “Addiction Is a Brain Disease, and It Matters.”¹³⁹ He argued that addiction is “a chronic, relapsing brain disorder,” and, therefore, a successful treatment would manage the illness but not be a “cure.” And he drew conclusions about both the criminal justice system (there should be less incarceration, more treatment) and society as a whole. “If the brain is the core of the problem, attending to the brain needs to be a core part of the solution.”¹⁴⁰

This vision of addiction as a brain disease is now over twenty-five years old and remains controversial.¹⁴¹ Not surprisingly, neuroscience research has probed

139. Alan I. Leshner, *Addiction Is a Brain Disease, and It Matters*, *Sci.*, Oct. 3, 1997, at 45, <https://doi.org/10.1126/science.278.5335.45>.

140. *Id.*

141. See, e.g., Neil Levy, *Addiction Is Not a Brain Disease (and It Matters)*, 4 *Frontiers Psychiatry* 24 (2013), <https://doi.org/10.3389/fpsy.2013.00024>; Markus Heilig et al., *Addiction as a Brain Disease Revised: Why It Still Matters, and the Need for Consilience*, 46 *Neuropsychopharmacology* 1715

many issues of addiction extensively, and much has been learned about pathways associated with addictive behavior, whether regarding illegal drugs, alcohol, or nicotine.¹⁴² Neuroscience has made progress in understanding how the brain acts during the three stages of addiction: binge/intoxication, withdrawal/negative affect, and preoccupation/anticipation. It has highlighted the important role of the neurotransmitter dopamine in many aspects of addiction. It has not, however, produced useful answers to predicting, preventing, or treating addiction, or to distinguishing between addiction and other kinds of uses of addictive substances. As is true with much of brain science, currently neuroscience can mainly tell us that “it is very complicated.”

That should change with increasing research, and, when it does, courts may be affected in at least two ways: “wholesale” and “retail.” The wholesale changes may be legal or policy shifts in dealing with addiction and addictive drugs, whether driven by legislatures, agencies, or the courts, similar to the changes in the constitutionality of punishment for some crimes committed by juveniles.

The U.S. Supreme Court long ago addressed some of the issues around addiction in two cases asking whether the Eighth Amendment’s prohibition on cruel and unusual punishment was violated by convictions for narcotics addiction (*Robinson v. California*, 370 U.S. 660 (1962)) or for being drunk in a public place (*Powell v. Texas*, 392 U.S. 514 (1968)). A 6–2 majority answered “yes” in *Robinson*; six years later, five justices answered “no” in *Powell*, although without being able to muster a majority opinion. As to crimes of addiction, *Robinson* has largely become a dead letter, but new neuroscience evidence might revive that doctrine.

For most judges, the issues of applying addiction neuroscience in cases will more likely be at the “retail” level. Strong neuroscience evidence that could distinguish “addiction” from lesser attachments might prove relevant in some cases, as could powerful evidence of a significant predisposition to addiction. New, well-proven treatments for addiction to various substances could spur changes in sentencing, parole, or probation conditions. Effective methods to block the effects of addictive substances, such as anti-drug vaccines that have been studied, might also affect individual cases. At this stage, these applications of

(2021), <https://doi.org/10.1038/s41386-020-00950-y>; John Davies, *Addiction Is Not a Brain Disease*, 26 *Addiction Res. & Theory* 1 (2017), <https://doi.org/10.1080/16066359.2017.1321741>.

142. See, e.g., George R. Uhl, George F. Koob & Jennifer Cable, *The Neurobiology of Addiction*, 1451 *Ann. N.Y. Acad. Scis.* 5 (2019), <https://doi.org/10.1111/nyas.13989>; Nora D. Volkow, Michael Michaelides & Ruben Baler, *The Neuroscience of Drug Reward and Addiction*, 99 *Physiological Rev.* 2115 (2019), <https://doi.org/10.1152/physrev.00014.2018>; Nora D. Volkow & Maureen Boyle, *Neuroscience of Addiction: Relevance to Prevention and Treatment*, 175 *Am. J. Psychiatry* 729 (2018), <https://doi.org/10.1176/appi.ajp.2018.17101174>. Koob has been director of the NIH’s National Institute on Alcohol Abuse and Alcoholism since 2014; Volkow (irrelevantly but interestingly a great-granddaughter of Leon Trotsky) has directed the NIH’s National Institute on Drug Abuse since 2003.

addiction neuroscience remain speculative; that might change before this manual reaches its fifth edition.

Conclusion

Atomic physicist Niels Bohr is often credited with having said, “It is always hard to predict things, especially the future.”¹⁴³ It seems highly likely that the massively increased understanding of the human brain that neuroscience is providing will have significant effects on the law and, more specifically, on the courts. Just what those effects will be cannot be accurately predicted, but we hope that this reference guide will provide some useful background to help judges cope with whatever neuroscience evidence comes their way.

143. This quotation has been attributed to many people, including Yogi Berra, but Bohr seems to be the most common nominee. Recently, one of us tracked down evidence that it actually was coined in the 1940s or perhaps 1930s by a Danish cartoonist, author, and inventor named Robert Storm Petersen. See *Robert Storm Petersen*, Wikipedia, <https://perma.cc/FR98-A6WA>. For a few more details, see Henry T. Greely, *The Death of Roe and the Future of Ex Vivo Embryos*, 9 J.L. & Biosciences 1, 2 & n.3 (2022), <https://doi.org/10.1093/jlb/lzac019>.

Glossary of Terms

axon. Portion of a neuron cell that transmits messages in the form of electrical impulses to dendrites of other neurons.

BOLD response (the blood-oxygen-level-dependent response). A measure used in functional MRI (fMRI) to allow an indirect measure of brain activity by examining how such activity affects blood flow.

brain stem. One of three major parts of the brain, joining the brain to the spinal cord and regulating autonomic functions such as heart rate and digestion, and, to an extent, the neural processing of the cerebellum.

central nervous system. The brain, the spinal cord, and other nerves directly connecting to the brain.

cerebellum. A brain structure at the top of the brain stem that keeps a library of learned motor skills.

cerebrum. Largest area of the brain (composed of two hemispheres) that controls cognitive functions.

computerized axial tomography (CAT scans). A multidimensional, computer-assisted X-ray machine that reveals the structural image of the brain in three dimensions.

corpus callosum. Axions that connect the two hemispheres of the brain.

cortex (cerebral cortex). Folded surface of the cerebrum that consists of gray matter and includes subcortical structures such as the thalamus, the hypothalamus, the basal ganglia, and the amygdala.

cortisol. A hormone produced by the adrenal glands that is released in response to stress.

deep brain stimulation (DBS). An electrical lead is implanted into a specific region of the brain and an electrical pulse is generated that affects the functioning of neurons in the nearby area.

dendrite. Portion of a neuron cell that receives messages from other neurons and relays them to the cell's nucleus.

diffusion tensor imaging (DTI). A technique that examines the way water diffuses through brain tissue to indicate the location of white matter.

electroencephalography (EEG). A technique that uses electrodes placed on the head to measure neural activity near the scalp.

endorphins. Hormones secreted by the pituitary gland in the brain that are associated with pain relief and a sense of well-being.

frontal lobe. Portion of the cerebrum near the forehead, associated with higher cognitive processes such as decision-making, reasoning, and planning.

functional near-infrared spectroscopy (fNIRS). A technique that detects changes in hemoglobin within the brain by measuring differences in the absorption of light that occurs through blood flow to different regions of the brain.

glial cells. Cells of the brain responsible for producing and maintaining the myelin sheaths that insulate axons and serve as a special immune system for the brain.

hippocampus. Area of the brain that is essential for making most kinds of memories and learning.

hypothalamus. A small structure at the base of the brain that regulates body temperature, hunger, thirst, and fatigue.

magnetic resonance imaging (MRI). The dominant neuroimaging technology, which uses a strong magnetic field to produce detailed images of the brain's structure and function.

myelin. Fatty substance encasing longer axons that helps insulate the axon and increases the strength and efficiency of the electrical signal.

neuron. Cells of the nervous system that transmit and receive nerve impulses from one such cell to another in a complex way that is responsible for brain function.

neurotransmitters. Chemical molecules that are released from the neuron into the synaptic cleft when a nerve impulse reaches the end of an axon. Some of the best-known neurotransmitters are dopamine, serotonin, glutamate, and acetylcholine.

nucleus accumbens. A small subcortical region in each hemisphere of the cerebrum that releases dopamine in response to rewarding experiences.

occipital lobe. Portion of the cerebrum at the back of the skull primarily concerned with vision.

olfactory bulb. Area of the brain that makes new neurons and plays a huge role in our senses of smell and taste.

oxytocin. A hormone that works as a neurotransmitter and has been linked with trust and bonding.

parietal lobe. Portion of the cerebrum at the top and toward the back of the skull that is concerned with reception and processing of sensory information from the body.

peripheral nervous system. Neural cells that are not part of the central nervous system.

positron emission tomography (PET scans). A brain imaging technique based on a radioactive tracer substance that measures the density of neural receptors.

single photon emission computed tomography (SPECT scans). A technique like PET scans using a different form of radioactive tracer.

substantia nigra. Area of the brain stem that generates the neurotransmitter dopamine as part of the brain's reward system.

synaptic cleft. The small space between two neurons where neurotransmitters are released.

temporal lobe. Portion at the sides of the cerebrum involved in hearing, language, memory storage, and emotion.

tesla (T). Measure of the strength of the magnetic field of an MRI scanner.

thalamus. A brain structure at the top of the brain stem that carries information to and from the cerebrum and other brain structures. This structure is especially important for transmission of sensory information such as vision, hearing, touch, and proprioception (one's sense of the position of the parts of one's body).

transcranial magnetic stimulation (TMS). A technique that generates a magnetic field near the skull that disrupts neural activity and alters brain function in a nearby region.

Reference Guide on Mental Health Evidence

KIRK HEILBRUN, DAVID DEMATTEO, AND PAUL S. APPELBAUM

Kirk Heilbrun, Ph.D., is Professor of Psychological and Brain Sciences, Department of Psychological and Brain Sciences, Drexel University.

David DeMatteo, J.D., Ph.D., is Professor of Psychological and Brain Sciences, Department of Psychological and Brain Sciences, and Professor of Law, Thomas R. Kline School of Law, Drexel University.

Paul S. Appelbaum, M.D., is the Elizabeth K. Dollard Professor of Psychiatry, Medicine, and Law, and Director, Center for Law, Ethics, and Psychiatry, Department of Psychiatry, Columbia University Irving Medical Center and New York State Psychiatric Institute.

CONTENTS

Overview of Mental Health Evidence, 1272

Range of Legal Cases in Which Mental Health Issues Arise, 1272

Retrospective, Contemporaneous, and Prospective Assessments, 1274

Diagnosis vs. Functional Impairment, 1277

Mental Health Experts, 1279

Psychiatrists, 1279

Psychologists, 1280

Other Mental Health Professionals, 1282

Differences Between Clinical and Forensic Contexts, 1283

Purpose, 1283

Client, 1284

Relationship, 1284

Voluntariness and Autonomy, 1284

Confidentiality, 1284

Consent, 1285

Response Style, 1285

Data Collection and Sources of Information, 1285

Pace, 1286

Setting, 1286

Report, 1286

Testimony, 1286

Diagnosis of Mental Disorders, 1287

Nomenclature and Typology: DSM-5 and DSM-5-TR, 1287

Major Diagnostic Categories, 1288

- Response Style, 1290
 - Malingering and Exaggeration, 1290
 - Other Response Styles, 1292
- Functional Impairment Due to Mental Disorders, 1292
 - Impact of Mental Disorders on Functional Capacities, 1292
 - Assessment of Functional Legal Capacities, 1293
 - Clinical Examination, 1294
 - Structured Assessment Techniques, 1295
- Predictive Assessments, 1297
 - Prediction of Violence, 1298
 - Approaches to Prediction of Violence, 1299
 - Limitations of Violence Prediction, 1301
 - Predictions of Future Functional Impairment, 1303
 - Risk–Need–Responsivity (RNR) Assessments, 1304
 - Appraising Risk, 1305
 - Appraising Need, 1305
 - Appraising Responsivity, 1306
- Treatment of Mental Disorders, 1306
 - Treatment with Medication, 1307
 - Psychological Treatments, 1307
 - Behavioral Therapy and Cognitive Behavioral Therapy, 1308
 - Specialized Treatments, 1308
 - Treatment of Functional Impairments, 1309
 - Prediction of Responses to Treatment, 1309
- Treatment to Reduce Risk of Reoffending, 1311
 - Treatment with Medications, 1311
 - Psychological Treatments, 1312
- Limitations of Mental Health Evidence, 1313
 - Ultimate–Issue Testimony, 1313
 - “Reasonable Degree of Certainty,” 1315
- Evaluating Evidence from Mental Health Experts, 1316
- What Are the Qualifications of the Expert?, 1316
 - Training, 1316
 - Experience, 1318
 - Licensure and Board Certification, 1320
 - Licensure, 1320
 - Board Certification, 1321
 - Prior Relationship with the Subject of the Evaluation, 1322
- How Was the Assessment Conducted?, 1324
 - Was the Evaluatee Examined in Person?, 1325
 - Did the Evaluatee Cooperate with the Assessment?, 1327
 - Was the Evaluation Conducted in Adequate Circumstances?, 1328
 - Were the Appropriate Records Reviewed?, 1329
 - Was Information Gathered from Collateral Informants?, 1331

Reference Guide on Mental Health Evidence

- Were Medical Diagnostic Tests Performed?, 1331
- Was the Evaluatee's Functional Impairment Assessed Directly?, 1332
- Was the Possibility of Malingering or Defensiveness Considered?, 1333
- Was a Structured Diagnostic or Functional Assessment Instrument
or Test Used?, 1334
 - Has the Reliability and Validity of the Instrument or
Test Been Established?, 1334
 - Does the Person Being Evaluated Resemble the Population
for Which the Instrument or Test Was Developed?, 1335
 - Was the Instrument or Test Used as Intended by Its Developers?, 1337
 - Training, 1337
 - Administration, 1338
 - Scoring, 1339
- How Was the Expert's Judgment Reached Regarding the Legally
Relevant Question?, 1339
 - Were the Findings of the Assessment Applied Appropriately
to the Question?, 1340
 - Were Diagnostic and Functional Issues Distinguished?, 1340
 - Were the Limitations of the Assessment and the Conclusions
Acknowledged?, 1341
 - Are Opinions Based on Valid Empirical Data Rather Than
Theoretical Formulations?, 1341
- Case Example, 1342
 - Facts of the Case, 1342
 - Testimony in the Federal Sentencing Hearing, 1343
 - Questions for Consideration: Dr. A, 1344
 - Questions for Consideration: Dr. B, 1345
- Glossary of Terms, 1347
- References on Mental Health Diagnosis and Treatment, 1351
- References on Mental Health and Law, 1351

Overview of Mental Health Evidence

Range of Legal Cases in Which Mental Health Issues Arise

Evidence presented by mental health experts is common to a broad array of legal cases—criminal and civil.¹ In the criminal realm, these include assessments of defendants' mental states at the time of their alleged offenses (e.g., criminal responsibility, diminished capacity²) and subsequent to the offenses, but prior to the initiation of the adjudicatory process (e.g., competence to consent to a search, capacity to waive *Miranda* rights³).⁴ As cases move toward adjudication, defendants' competence to stand trial or to represent themselves at trial may need to be evaluated.⁵ Postconviction, mental health evidence may be introduced for sentencing, including suitability for probation and conditions of probation.⁶

1. See Eric Y. Drogin et al., *Handbook of Forensic Assessment: Psychological and Psychiatric Perspectives* (2011) (discussing role of mental health evidence in range of criminal and civil cases); Gary B. Melton et al., *Psychological Evaluations for the Courts: A Handbook for Mental Health Professionals and Lawyers* (4th ed. 2018) (discussing range of criminal and civil cases in which mental health evidence may be informative); Christopher Slobogin, Thomas L. Hafemeister & Douglas Mossman, *Law and the Mental Health System: Civil and Criminal Aspects* (7th ed. 2020) (discussing use of mental health evidence in criminal and civil contexts). This reference guide will focus on adult defendants and litigants and will not discuss matters pertaining to the evaluation of juveniles. For a review of relevant case law in this area, see Melton et al., *supra*. See also Kirk Heilbrun et al., *Evaluating Juvenile Transfer and Disposition: Law, Science, and Practice* (2017).

2. 18 U.S.C. § 17 (defining standard and burden of proof for insanity defense); *Clark v. Arizona*, 548 U.S. 735 (2006) (ruling on the use of testimony for diminished capacity).

3. See Thomas Grisso, *Evaluating Competencies: Forensic Assessments and Instruments* (2d ed. 2003); *Colorado v. Connelly*, 479 U.S. 157 (1986) (holding that a mental health condition alone will not make a confession involuntary under the Fourth Amendment, but may be used as a factor in assessing voluntariness of defendant's confession); *Miranda v. Arizona*, 384 U.S. 436 (1966) (holding confessions inadmissible unless suspect is made aware of rights and waives them); *United States v. Elrod*, 441 F.2d 353 (5th Cir. 1971) (holding that a person of below-average intelligence may be deemed incapable of giving consent). See also Wayne R. LaFare, *Search and Seizure: A Treatise on the Fourth Amendment* 92–93 (2004); Wayne R. LaFare, Jerold H. Israel & Nancy J. King, *Criminal Procedure* 363–65 (4th ed. 2004); Brian S. Love, *Beyond Police Conduct: Analyzing Voluntary Consent to Warrantless Searches by the Mentally Ill and Disabled*, 48 St. Louis U.L.J. 1469 (2004).

4. See generally Stephen J. Morse, *Mental Disorder and Criminal Law*, 101 J. Crim. L. & Criminology 885 (2011) (discussing use of mental health evidence in criminal-law contexts).

5. *Indiana v. Edwards*, 554 U.S. 164 (2008) (finding that the standards for competency to stand trial and to represent oneself need not be the same); *Farretta v. California*, 422 U.S. 806 (1975) (upholding defendant's right to refuse counsel and represent himself); *Pate v. Robinson*, 383 U.S. 375 (1966) (holding that the Due Process Clause of the Fourteenth Amendment does not allow a mentally incompetent criminal defendant to stand trial); *Dusky v. United States*, 362 U.S. 402 (1960) (establishing standard for competence to stand trial).

6. Roger W. Haines Jr., Frank O. Bowman & Jennifer C. Woll, *Federal Sentencing Guidelines Handbook: Text and Analysis* §§ 5B1.3(d)(5), 5D1.3(d)(5), 5H1.3 (2007–2008).

Capital cases may raise unique questions regarding a condemned prisoner's competence to waive appeals or to be executed.⁷ Postconfinement, mental health considerations may enter into parole determinations.

Mental health evidence in civil litigation is frequently introduced in personal-injury cases, where emotional harms may be alleged with or without concomitant physical injury.⁸ Issues of contract may turn on the competence of a party at the time that the contract was concluded or whether that person was subject to undue influence,⁹ and similar questions may be at the heart of litigation over wills and gifts.¹⁰ Broader questions of competence to conduct one's affairs are considered in guardianship cases,¹¹ and more esoteric ones may arise in litigation challenging a person's competence to enter into a marriage or to vote.¹²

7. See *Panetti v. Quarterman*, 551 U.S. 930 (2007) (holding that defendants sentenced to death must be competent at the time of execution); *Atkins v. Virginia*, 536 U.S. 304 (2002) (finding that executing intellectually disabled defendants constitutes cruel and unusual punishment under the Eighth Amendment); *Stewart v. Martinez-Villareal*, 523 U.S. 637 (1998) (holding that death row prisoners are not barred from filing incompetence-to-be-executed claims by dismissal of previous federal habeas petitions); *Ford v. Wainwright*, 477 U.S. 399 (1986) (upholding the common-law bar against executing incompetent inmates and holding that an inmate is entitled to a judicial hearing before being executed); *Rees v. Peyton*, 384 U.S. 312 (1966) (formulating the test for competency to waive further proceedings as requiring that the petitioner "appreciate his position and make a rational choice with respect to continuing or abandoning further litigation or on the other hand whether he is suffering from a mental disease, disorder, or defect which may substantially affect his capacity in the premises"). See also David DeMatteo et al., *Forensic Mental Health Assessments in Death Penalty Cases* (2011) (discussing role of mental health evidence in all phases of capital cases).

8. *Sheely v. MRI Radiology Network, P.A.*, 505 F.3d 1173 (11th Cir. 2007) (holding that damages are available under § 504 of the Rehabilitation Act when emotional distress was foreseeable); *Cooper v. FAA*, No. 07-1383 (N.D. Cal. Aug. 2008), *rev'd and remanded*, 596 F.3d 538 (9th Cir. 2010) (discussing mental distress as a result of disclosure of personal information); *Albright v. United States*, 732 F.2d 181 (C.A.D.C. 1984) (holding that alleging mental distress is sufficient to confer standing); *Molien v. Kaiser Found. Hosps.*, 616 P.2d 813 (Cal. 1980) (holding that plaintiff who is direct victim of negligent act need not be present when act occurs to recover for subsequent emotional distress); *Dillon v. Legg*, 441 P.2d 912 (Cal. 1968) (allowing recovery based on emotional distress not accompanied by physical injury); *Roes v. FHP, Inc.*, 985 P.2d 661 (Haw. 1999) (allowing assessment of damages for negligent infliction of emotional distress when plaintiff was in actual physical peril, even if no injury was suffered); *Rodriguez v. State*, 472 P.2d 509 (Haw. 1970) (permitting recovery where a reasonable person would suffer serious mental distress as a result of defendant's behavior).

9. See generally E. Allan Farnsworth, *Contracts* 228-33 (2004); John Parry & Eric Y. Drogin, *Mental Disability Law, Evidence, and Testimony* 151-52, 185-86 (2007).

10. See generally William M. McGovern Jr. & Sheldon F. Kurtz, *Wills, Trusts and Estates Including Taxation and Future Interests* 292-99 (6th ed. 2021); Parry & Drogin, *supra* note 9, at 149-51, 182-85.

11. Parry & Drogin, *supra* note 9, at 138-47, 177-81.

12. *Id.* at 54. See *Missouri Prot. & Advoc. Servs. v. Carnahan*, 499 F.3d 803 (8th Cir. 2007) (upholding a state law allowing disenfranchisement of persons under guardianship because it permits individualized determinations of capacity to vote); *Doe v. Rowe*, 156 F. Supp. 2d 35

Suits alleging infringement of the statutory and constitutional rights of persons with mental disorders (e.g., under the Americans with Disabilities Act (ADA) or the Civil Rights of Institutionalized Persons Act) often involve detailed consideration of psychiatric diagnosis and treatment and of institutional conditions.¹³ Allegations of professional malpractice by mental health professionals, including failure to protect foreseeable victims of a patient's violence,¹⁴ invariably call for mental health expert testimony, as do commitment proceedings for the hospitalization of persons with mental disorders¹⁵ or who are alleged to be dangerous sex offenders.¹⁶

Retrospective, Contemporaneous, and Prospective Assessments

Retrospective assessments are called for when criminal defendants assert insanity or diminished-capacity defenses, claiming that their state of mind at the time of

(D. Me. 2001) (finding a state law denying the vote to anyone under guardianship by reason of mental disability in violation of the Equal Protection Clause of the U.S. Constitution and Title II of the Americans with Disabilities Act (ADA)).

13. *Pennsylvania Dep't of Corr. v. Yeskey*, 524 U.S. 206 (1998) (holding that ADA coverage extended to prisoners); *McAlinden v. Cnty. of San Diego*, 192 F.3d 1226 (9th Cir. 1999), *cert. denied*, 120 S. Ct. 2689 (2000) (reversing summary judgment against plaintiff who alleged that anxiety and somatoform disorders impaired major life activities of sexual relations and sleep); *Gates v. Cook*, 376 F.3d 323 (5th Cir. 2004) (upholding district court's finding that prison conditions, including inadequate mental health provisions, violated the Eighth Amendment of the U.S. Constitution); *Steele v. Thiokol Corp.*, 241 F.3d 1248 (10th Cir. 2001) (finding that major life activity under the ADA, involving interacting with others, not substantially impaired by obsessive-compulsive disorder); *Anderson v. North Dakota State Hosp.*, 232 F.3d 634 (8th Cir. 2000) (finding that a plaintiff's fear of snakes did not limit ability to work); *Sinkler v. Midwest Prop. Mgmt.*, 209 F.3d 678 (7th Cir. 2000) (holding driving phobia did not substantially limit major life activity of working and hence was not an impairment under the ADA); *Clark v. State of Cal.*, 123 F.3d 1267 (9th Cir. 1997) (finding state not immune on Eleventh Amendment grounds to suit alleging discrimination under ADA by developmentally disabled inmates); *Gaul v. AT&T, Inc.*, 955 F. Supp. 346 (D.N.J. 1997) (finding that depression and anxiety disorders may constitute a mental disability under the ADA).

14. *Tarasoff v. Regents of the Univ. of Cal.*, 551 P.2d 334 (Cal. 1976).

15. *Addington v. Texas*, 441 U.S. 418 (1979) (holding that standard of proof for involuntary commitment is clear and convincing evidence); *O'Connor v. Donaldson*, 422 U.S. 563 (1975) (holding unconstitutional the confinement of a nondangerous mentally ill person capable of surviving safely in freedom alone or with assistance); *Lake v. Cameron*, 364 F.2d 657 (D.C. Cir. 1966) (discussing least restrictive alternative doctrine as applied to institutionalized individuals).

16. *United States v. Comstock*, 560 U.S. 126 (2010); *Kansas v. Crane*, 534 U.S. 407 (2002); *Kansas v. Hendricks*, 521 U.S. 346 (1997). See David DeMatteo et al., *A National Survey of United States Sexually Violent Predator Legislation: Policy, Procedures, and Practice*, 14 Int'l J. Forensic Mental Health 245 (2015), <https://doi.org/10.1080/14999013.2015.1110847> (discussing sexually violent predator legislation in all U.S. jurisdictions).

the crime should excuse or mitigate the consequences of their behaviors,¹⁷ or when questions are raised about competence at some point in the past to waive legal rights (e.g., waiver of *Miranda* rights).¹⁸ In civil contexts, challenges to the capacity of a now-deceased testator to write a will or of a party to enter into a contract, among other issues, will call for a similar look back at a person's functioning at some point in the past.¹⁹ A variety of sources of information are available for such assessments. In some cases (e.g., in criminal proceedings) the defendant is likely to be available for clinical examination, whereas in other cases the defendant will not be able to be assessed directly (e.g., challenges to a will). Although persons being evaluated will usually have an interest in portraying themselves in a particular light, direct assessments can still be valuable in determining the consistency of the reported symptoms with other aspects of a person's history and current status.²⁰ Whether or not a person can be assessed directly, information, including direct reports and contemporaneous records, from those who were in contact with the person before and during the time in question is usually an essential part of the evaluation. Sometimes the available data from all of these sources are so limited or contradictory that they will not allow an evaluator to reach an opinion on a person's state of mind at a point in the past. However, when evaluators are thorough, seeking information from multiple sources in relevant areas, they are usually able to provide the decision-maker with useful information.

Current-state evaluations are the most straightforward task for a mental health professional. In criminal-justice settings, concerns about a person's current competence to exercise or waive rights will call for such evaluations (e.g., competence to stand trial or competence to represent oneself at trial).²¹ Civil

17. See Ira K. Packer, *Evaluation of Criminal Responsibility* (2009) (discussing role of mental health professionals in evaluating a defendant's mental state at the time of the offense); see also David DeMatteo et al., *The United States Supreme Court's Enduring Misunderstanding of Insanity*, 52 N.M. L. Rev. 34 (2022) (presenting analysis of Supreme Court cases that illustrate the Court's confusion regarding the insanity defense).

18. *Retrospective Assessment of Mental States in Litigation: Predicting the Past* (Robert I. Simon & Daniel W. Shuman eds., 2002); Bruce Frumkin & Alfredo Garcia, *Psychological Evaluations and Competency to Waive Miranda Rights*, 9 *The Champion* 12 (2003); Alan Goldstein & Naomi E. Sevin Goldstein, *Evaluating Capacity to Waive Miranda Rights* (2010).

19. See Thomas G. Gutheil, *Common Pitfalls in the Evaluation of Testamentary Capacity*, 35 *J. Am. Acad. Psychiatry & L.* 514 (2007); Farnsworth, *supra* note 9, at 228–33; Kenneth I. Shulman et al., *Assessment of Testamentary Capacity and Vulnerability to Undue Influence*, 164 *Am. J. Psychiatry* 722 (2007), <https://doi.org/10.1176/ajp.2007.164.5.722>.

20. See Richard Rogers & Scott D. Bender, *Clinical Assessment of Malingering and Deception* (4th ed. 2018) (discussing importance of assessing response style in forensic mental health assessments).

21. See *Dusky v. United States*, 362 U.S. 402 (1960) (holding that a criminal defendant must understand the charges and be able to participate in his defense); *Godinez v. Moran*, 509 U.S. 389 (1993) (holding that a defendant competent to stand trial is also sufficiently competent to plead guilty or waive the right to legal counsel).

issues calling for contemporaneous assessments include workers' compensation and other disability claims and litigation alleging emotional harms due to negligent or intentional torts, workplace discrimination, and other harm-inducing situations.²² Careful consideration needs to be given to the possibility that persons being assessed in these circumstances could deliberately exaggerate or minimize the severity of their experienced symptoms to secure a more favorable outcome in their case.²³

The evaluation of a person's future mental state and consequent behaviors, by contrast, is particularly difficult, especially when the outcome being predicted is relatively infrequent.²⁴ In criminal proceedings, predictive assessments like this may come into play when bail is set,²⁵ at sentencing,²⁶ or as part of probation and parole decisions.²⁷ Predictive assessments often involve estimating the probable effectiveness of treatment—especially in the juvenile justice system, where a key consideration in decisions whether to transfer a juvenile to adult courts is often whether the juvenile is amenable to mental health treatment.²⁸ Predictions of behavior related to mental health disorders are also seen in civil cases, for example, in the civil commitments of persons with mental disorders and in statutes authorizing the commitment of dangerous sex offenders.²⁹ Damage assessments in civil cases that allege emotional harms will usually call for some estimate on the duration of symptoms and response to treatment.³⁰ The

22. See, e.g., *Rivera v. City of New York*, 392 F. Supp. 2d 644 (S.D.N.Y. 2005); *Quigley v. Barnhart*, 224 F. Supp. 2d 357 (D. Mass. 2002); *Kent v. Apfel*, 75 F. Supp. 2d 1170 (D. Kan. 1999); *Lahr v. Fulbright & Jaworski, L.L.P.*, 164 F.R.D. 204 (N.D. Tex. 1996); see also Jane Goodman-Delahanty & William E. Foote, *Evaluation for Workplace Discrimination and Harassment* (2011).

23. See *United States v. Binion*, 132 F. App'x 89 (8th Cir. 2005) (upholding an obstruction-of-justice conviction and sentencing determination based on a finding that defendant had feigned mental illness). See discussion in section titled "Response Style" below.

24. Joseph M. Livermore, Carl P. Malmquist & Paul E. Meehl, *On the Justifications for Civil Commitment*, 117 U. Pa. L. Rev. 75–96 (1968).

25. *United States v. Salerno*, 481 U.S. 739 (1987); *United States v. Farris*, No. 2:08cr145, 2008 WL 1944131 (W.D. Pa. May 1, 2008).

26. Tex. Code Crim. Proc. Ann. art. 37.071 (Vernon 1981); *Barefoot v. Estelle*, 463 U.S. 880 (1983).

27. See 28 C.F.R. § 2.19 (2008) for parole determination factors. For probation determination factors, see 18 U.S.C. § 3563 (2008). See generally Neil P. Cohen, *The Law of Probation and Parole* §§ 2, 3 (2008).

28. Michael G. Kalogerakis, *Handbook of Psychiatric Practice in Juvenile Court* 79–85 (1992).

29. See *O'Connor v. Donaldson*, 422 U.S. 563 (1975) (finding that a state may not confine a citizen who is nondangerous and capable of living by herself or with aid). For an example of a sex-offender civil-commitment statute, see Minn. Stat. § 253B.185 (2008). The constitutionality of civil commitment for dangerous sex offenders was upheld in *Kansas v. Hendricks*, 521 U.S. 346 (1997) (setting forth the procedures for the commitment of convicted sex offenders deemed dangerous because of a mental abnormality).

30. See Melton et al., *supra* note 1.

uncertainties of the course of mental disorders and their responsiveness to interventions are relevant—but so are the unknowable contingencies of life (e.g., what occurs with family, job, and friends). At best, predictive assessments can provide general statements of the probability of particular outcomes, with an acknowledgment of the uncertainties involved.³¹

Diagnosis vs. Functional Impairment

A diagnosis of mental disorder per se will almost never settle the legal question in a case in which mental health evidence is presented. However, a diagnosis may play a role in determining whether a claim or proceeding can go forward. With the insanity defense, for example, the impairments of understanding, appreciation, and behavioral control that constitute the various legal standards must be based, in one popular formulation, on a “mental disease or defect.”³² In the absence of a diagnosis of mental disorder (including intellectual disability or consequences of injury to the brain), an affirmative defense of insanity will not prevail.³³ Comparable situations exist in civil commitment proceedings and work-disability determinations.³⁴

Notwithstanding the threshold role played by a mental disorder diagnosis in many cases, the ultimate legal issue usually will turn on the impact of the mental disorder on the person’s functional abilities.³⁵ Those abilities may refer to the person’s cognitive capacities, including capacities to make legally relevant decisions, or *decisional capacities* (e.g., granting consent for the police to conduct a warrantless search, altering a will), or *performative capacities*, that is, the capacity to behave in a particular way (e.g., conforming one’s conduct to the requirements of the law, cooperating with an attorney in one’s own defense, resisting undue influence), or both (e.g., skill as a parent, competence to proceed with criminal

31. For a more detailed discussion of predictive assessment regarding future dangerousness, see section titled “Prediction of Violence” below.

32. The American Law Institute’s standard for the insanity defense reads, “A person is not responsible for criminal conduct if at the time of such conduct as a result of mental disease or defect he lacks substantial capacity either to appreciate the criminality of his conduct or to conform his conduct to the requirements of the law.” Model Penal Code and Commentaries § 4.01(1) (Official Draft and Revised Comments 1985) (adopted by American Law Institute, May 24, 1962). The federal insanity defense was codified in the Insanity Defense Reform Act of 1984, 18 U.S.C. § 17. See also *Durham v. United States*, 214 F.2d 862 (D.C. Cir. 1954) (“[A]n accused is not criminally responsible if his unlawful act was the product of mental disease or defect.”); note *United States v. Brawner*, 471 F.2d 969 (1972) (which overturned the *Durham* Rule or “product test”).

33. *Tennard v. Dretke*, 542 U.S. 274 (2004); *Bigby v. Dretke*, 402 F.3d 551 (5th Cir. 2005).

34. *Addington v. Texas*, 441 U.S. 418 (1979) (setting the burden of proof required for involuntary civil commitment as requiring clear and convincing evidence); see Social Security Administration Listing of Impairments, <https://perma.cc/5TNU-5JAD>.

35. Grisso, *supra* note 3.

adjudication). Many of the legal questions related to mental health evidence will involve a determination of the influence of a mental state or disorder on one or both of these sets of capacities. The mere presence of a mental disorder will almost always be insufficient for that purpose. Mental disorder in a criminal defendant, for example, if it does not interfere substantially with competence to stand trial, does not present a basis for postponing adjudication of the case.³⁶ Some degree of mental disorder, including neurocognitive impairment, without affecting relevant abilities, does not provide grounds for voiding a will.³⁷ This can be generalized to all criminal and civil competency determinations, most assessments of emotional harms, and probably to the majority of cases in which mental health testimony is offered: Unless a mental disorder can be shown to have affected a person's functional decisional or performative capacity, a diagnosis of mental disorder per se will not be determinative of the outcome.³⁸

Despite its importance to the adjudicative process, mental health evidence is often introduced in the context of a serious stigma that attaches to mental disorders³⁹ and considerable popular confusion regarding their nature, consequences, and susceptibility to treatment.⁴⁰ Diagnoses of mental disorders often are perceived to be less reliable and more subjective than diagnoses of other medical conditions.⁴¹ Symptoms of mental disorders may be seen as reflections of moral weakness or lack of will, and the impact of disorders on functional abilities may not be recognized or may occasionally be exaggerated.⁴² The potential impact and limits of current treatments are not widely understood. Even the various types of mental health professionals are frequently confused.⁴³ The first section of this reference guide clarifies these issues; subsequently, we address questions

36. *United States v. Valtierra*, 467 F.2d 125 (9th Cir. 1972); *United States v. Passman*, 455 F. Supp. 794 (D.D.C. 1978).

37. *Rossi v. Fletcher*, 418 F.2d 1169 (D.C. Cir. 1969); *In re Estate of Buchanan*, 245 A.D.2d 642 (N.Y. App. Div. 1997).

38. For a brief overview of competency evaluations, see Patricia A. Zapf & Ronald Roesch, *Mental Competency Evaluations: Guidelines for Judges and Attorneys*, 37 Ct. Rev. 28 (2000). For the underlying standard for competency to stand trial, see *Dusky v. United States*, 362 U.S. 402 (1960).

39. Bruce G. Link et al., *Measuring Mental Illness Stigma*, 30 Schizophrenia Bull. 511 (2004), <https://doi.org/10.1093/oxfordjournals.schbul.a007098>.

40. Bruce G. Link et al., *Stigma and Coercion in the Context of Outpatient Treatment for People with Mental Illnesses*, 67 Soc. Sci. & Med. 409 (2008), <https://doi.org/10.1016/j.socscimed.2008.03.015>.

41. Thomas A. Widiger, *Values, Politics, and Science in the Construction of the DSMs*, in *Descriptions and Prescriptions: Values, Mental Disorders, and the DSMs* 25 (John Z. Sadler ed., 2002).

42. Michael L. Perlin, "You Have Discussed Lepers and Crooks": *Sanism in Clinical Teaching*, 9 Clinical L. Rev. 683 (2003); Michael L. Perlin, "Half-Wracked Prejudice Leaped Forth": *Sanism, Pretextuality, and Why and How Mental Disability Law Developed as It Did*, 10 J. Contemp. Legal Issues 3 (1999); Michael L. Perlin, *The Hidden Prejudice: Mental Disability on Trial* (2000).

43. The degree of popular confusion is underscored by the results of a web-based search for "psychiatrist vs. psychologist," which turns up a remarkably large number of websites attempting to explain the differences between the two professions.

specifically related to the introduction of evidence by mental health experts and provide a case example.

Mental Health Experts

Evidence related to mental state and mental health disorders may be presented by experts from a number of disciplines, but it is most commonly introduced by psychiatrists or psychologists.

Psychiatrists

Psychiatrists are physicians who specialize in the diagnosis and treatment of mental disorders.⁴⁴ After college, they complete four years of medical school, during which they spend approximately two years in preclinical studies (e.g., physiology, pharmacology, genetics, pathophysiology), followed by two years of clinical rotations in hospital and clinic settings (e.g., medicine, surgery, pediatrics, obstetrics/gynecology, orthopedics, psychiatry).⁴⁵ Graduating medical students who elect to specialize in psychiatry enter residency programs of at least four years.⁴⁶ Accredited residencies must currently offer at least four months in a primary-care setting in internal medicine, family medicine, or pediatrics and at least two months of training in neurology.⁴⁷

After completing four years of residency training, a psychiatrist is considered board eligible, that is, able to take the certification examination of the American Board of Psychiatry and Neurology in adult psychiatry.⁴⁸ After successful completion of this examination process, the psychiatrist is designated board certified. Psychiatrists who desire more intensive training in a subspecialty area of psychiatry—for example, child and adolescent or addiction psychiatry—can complete a one- or two-year fellowship working in that subspecialty. A psychiatrist who has completed

44. Narriman C. Shahrokh & Robert E. Hales, *American Psychiatric Glossary* 157 (2003).

45. Medical schools in the United States are accredited by the Liaison Committee on Medical Education, which establishes general curricular and other standards that all schools must meet. Standards are available at <https://perma.cc/9KT4-LED7>. Students can elect to extend their medical-school training by taking additional time to conduct research or to obtain complementary training (e.g., in public health).

46. Residents who choose to combine adult- and child-psychiatry training can do so in a five-year program, or can follow their four years of adult residency with two years of child training. Some residents will also extend their residency training by adding a year or more during which they conduct laboratory or clinical research.

47. Psychiatric residencies are accredited by the Accreditation Council on Graduate Medical Education. Program requirements are available at <https://perma.cc/QX6X-ZP6H>.

48. Information on qualifications for board certification and the examination process is available from the American Board of Psychiatry and Neurology at <https://perma.cc/EN4Y-QFA6>.

an accredited fellowship⁴⁹ is eligible for additional board certification in that subspecialty.⁵⁰

Forensic psychiatry is a subspecialty that focuses on the interrelationships between psychiatry and the law.⁵¹ It is the forensic psychiatrist who is particularly likely to offer evidence as part of legal proceedings. Fellowship training in forensic psychiatry involves a one-year program in which fellows are taught forensic evaluation for civil and criminal litigation and become involved in the treatment of persons with mental disorders in the correctional system.⁵²

Psychologists

Psychologists receive graduate training in the study of mental health processes and behavior.⁵³ Only a subset of psychologists—those trained in applied specialty areas—evaluate and treat persons with psychological or behavioral problems, including clinical, counseling, health, neuro-, rehabilitation, and school psychologists. Independent practice in applied specialty areas of psychology requires licensure from the appropriate state licensure board and generally requires a doctoral degree and postgraduate clinical experience. Although use of the term *psychologist* is restricted in many jurisdictions to licensed psychologists,⁵⁴

49. Accredited subspecialty training is currently available in addiction, child and adolescent, consultation-liaison, forensic, and geriatric psychiatry. Psychiatrists are also eligible for training in brain-injury medicine, hospice and palliative medicine, pain medicine, and sleep medicine. See subspecialty certification standards at <https://perma.cc/UB3S-7TCH>. Fellowship programs also exist in some subspecialty areas for which accreditation and board certification are not available (e.g., research, psychopharmacology, and public and community psychiatry).

50. Typically, when new subspecialties are recognized and accreditation standards are developed, a certain period of time (e.g., five years) is allowed for psychiatrists who have gained expertise in that area by virtue of experience or alternative training to achieve board certification; many psychiatrists who are today board certified in a subspecialty have not completed a fellowship.

51. See the definition of forensic psychiatry offered by the American Academy of Psychiatry and the Law: “Forensic psychiatry is a medical subspecialty that includes research and clinical practice in the many areas in which psychiatry is applied to legal issues.” <https://perma.cc/VPP2-22JF>. Psychiatrists who have been certified in adult or child psychiatry by the American Board of Psychiatry and Neurology, and who have completed a forensic psychiatry fellowship, can take the examination for subspecialty certification in forensic psychiatry. A description of the requirements for certification can be found at <https://perma.cc/T74N-N5NF>. Board certification must be renewed by taking a recertification examination every ten years or participating in an alternative maintenance of certification program, see details at <https://perma.cc/K3LR-33XM>.

52. See the accreditation standards in forensic psychiatry at <https://perma.cc/DW8U-EX6B>.

53. The American Psychological Association defines the field of psychology in this way: “Psychology is the study of the mind and behavior. The discipline embraces all aspects of the human experience—from the functions of the brain to the actions of nations, from child development to care for the aged. In every conceivable setting from scientific research centers to mental health care services, ‘the understanding of behavior’ is the enterprise of psychologists.” <https://perma.cc/PFV3-MJHQ>.

54. See, e.g., Mass. Gen. Laws ch. 112, § 122; N.Y. Educ. Law § 7601.

the term may be applied to those in academia or to those with master's-level training in psychology in certain contexts.⁵⁵

Those pursuing doctoral-level training as psychologists must complete a graduate program that typically takes four to six years. Those who intend to provide clinical services generally receive training in clinical psychology, counseling psychology, neuropsychology, or school psychology,⁵⁶ involving three to five years of graduate study plus a year of clinical internship.⁵⁷ Psychology graduate programs award either the Ph.D. or Psy.D. (professional psychology) degree.⁵⁸ Ph.D. programs that adhere to the scientist-practitioner training model emphasize both research and clinical training, and students are required to gain clinical experience and complete a research-based dissertation. Psy.D. programs ordinarily focus on clinical training and have less rigorous research requirements.⁵⁹

Postdoctoral fellowships have been developed in forensic psychology, generally involving a one-year program with didactic and clinical training in forensic evaluation.⁶⁰ Board certification is offered by several entities, but the most established is the American Board of Professional Psychology (ABPP),⁶¹ which provides certification in fifteen specialty areas. The American Board of Forensic Psychology (ABFP), one of the ABPP's specialty boards, provides board certification in forensic psychology. Board certification in forensic psychology through

55. Note that the American Psychological Association urges that the term be restricted to persons with doctoral degrees in psychology: "Psychologists have a doctoral degree in psychology from an organized, sequential program in a regionally accredited university or professional school . . . it is [the] general pattern to refer to master's-level positions as counselors, specialists, clinicians, and so forth (rather than as 'psychologists')." <https://perma.cc/S38C-54YG>.

56. Other psychology programs offer training in experimental, social, and cognitive psychology, for example, with the intent of producing graduates who will pursue research or teaching careers but will not engage in clinical work. *United States v. Fishman*, 743 F. Supp. 713, 723 (N.D. Cal. 1990) (excluding the expert testimony of a social psychologist holding a Ph.D. in sociology).

57. Accreditation of programs in clinical, counseling, and school psychology is undertaken by the Commission on Accreditation of the American Psychological Association. Accreditation standards are available at <https://perma.cc/Q3YM-5U9Y>.

58. See sample Ph.D. program curricula for programs at University of Illinois at Urbana-Champaign, <https://perma.cc/MBL5-8TX9>; Indiana University, <https://perma.cc/82JY-66L3>; and University of California at Los Angeles, <https://perma.cc/MN2Q-KFPJ>. See sample Psy.D. curricula for programs at William James College, <https://perma.cc/6CQ7-FNHW>; and Wisconsin School of Professional Psychology, <https://perma.cc/LZ4F-6G5P>.

59. For a discussion of the so-called Vail model on which Psy.D. training is based, see John C. Norcross & Patricia H. Castle, *Appreciating the PsyD: The Facts*, 7 *Eye on Psi Chi* 22 (2002), <https://doi.org/10.24839/1092-0803.Eye7.1.22>.

60. See, e.g., Description of Program at University of Massachusetts Medical School, <https://perma.cc/YJD4-44Q4>.

61. See American Board of Professional Psychology website, <https://perma.cc/D2ES-YCRS>. ABPP is the oldest and best-established specialty certification organization for psychologists, with the most rigorous process for becoming board certified.

ABPP/ABFP is a rigorous process that includes a credential review documenting sufficient education, training, and experience; completion of a written examination; submission and review of forensic practice samples; and a three-hour oral examination.⁶²

Other Mental Health Professionals

Professionals with a variety of other types of training also provide mental health services, including services that are generally referred to as psychotherapy or counseling with individuals, couples, or groups. The best established of these professions is social work. Schools of social work offer two-year programs that lead to a master's degree (MSW), and students can elect a track that is often referred to as psychiatric social work, which involves instruction and experience in psychotherapy.⁶³ Graduate social workers can obtain state licensure after they complete a period of supervised practice and an examination, after which they become a licensed independent clinical social worker (LICSW), with variation in nomenclature across the states.⁶⁴ Social workers may offer psychotherapeutic or counseling services through social service agencies or in private practice. Recently, a subspecialty of forensic social work has begun to develop, involving social workers with experience in the criminal justice system.⁶⁵

Although psychiatrists and doctoral-level psychologists generally provide expert evidence related to mental health issues, courts will sometimes admit testimony from other mental health professionals.⁶⁶ Given that training and

62. The requirements for board certification in forensic psychology through ABPP/ABFP can be found at <https://perma.cc/PY8U-RBQD>.

63. See, e.g., curricula for social-work training at Columbia School of Social Work (<https://perma.cc/A5UP-NYQB>) and Smith College (<https://perma.cc/XGY2-N8UB>).

64. The Association of Social Work Boards provides an overview of state licensure requirements at <https://perma.cc/Z886-M3XV>.

65. See the National Association of Forensic Social Work's description of forensic social work at <https://perma.cc/6ZDE-RL3H>. Postgraduate certification programs for forensic social workers are also being developed (e.g., University of Nevada at Las Vegas, <https://perma.cc/8Y2F-AMZX>).

66. *Leblanc v. Coastal Mech. Servs., LLC*, No. 04-80611-CIV, 2005 WL 5955027 (S.D. Fla. Sept. 7, 2005) (finding that a marriage and family counselor holding a Ph.D. in family therapy and bachelor's and master's degrees in psychology, and having a record of relevant publications, may be qualified to offer helpful testimony about a plaintiff's alleged psychological condition); *Jenkins v. United States*, 307 F.2d 637 (D.C. Cir. 1962) ("The critical factor in respect to admissibility is the actual experience of the witness and the probable probative value of his opinion. . . . The determination of a psychologist's competence to render an expert opinion based on his findings as to the presence or absence of mental disease or defect must depend upon the nature and extent of his knowledge. It does not depend upon his claim to the title 'psychologist.'"); *United States v. Azure*, 801 F.2d 336, 342 (8th Cir. 1986) ("The social worker was most likely qualified as an expert under Rule 702. . . ."); see also *United States v. Raya*, 45 M.J. 251 (1996) (finding that trial court's admission of expert testimony from a social worker on whether the victim suffered from post-traumatic

experience vary considerably, and titles may be used inconsistently, courts usually require an individualized inquiry into the qualifications of the proposed expert.

Differences Between Clinical and Forensic Contexts

The view of a psychologist or psychiatrist as a helping professional in a traditional clinical context can be contrasted with their roles in forensic contexts. Several of the key differences between clinical assessments and forensic assessments are described below.⁶⁷

Purpose

The primary distinction between traditional clinical assessments and forensic assessments is the different purposes of the evaluations. In a clinical context, the purpose of an assessment is to diagnose mental health problems and develop a treatment plan; the purpose of a forensic assessment is to assist a legal decision-maker in making a decision about a criminal offender or civil litigant. Ethical principles governing clinical practice emphasize benefiting the patient and avoiding harm, while the ethics of forensic evaluation center on a truthful

stress disorder (PTSD) was not an abuse of discretion) *and* United States v. Johnson, 35 M.J. 17 (1992) (holding social worker qualified to render opinion that child suffered trauma). Note, however, that not all courts have been receptive to social-worker testimony offered as expert opinion on the diagnosis of PTSD, e.g., *Blackshear v. Werner Enters., Inc.*, No. 2004-4—WOB, 2005 WL 6011291 (E.D. Ky. May 19, 2005); *Neely v. Miller Brewing Co.*, 246 F. Supp. 2d 866 (S.D. Ohio 2003). For more restrictive approaches to testimony by non-Ph.D. psychologists, *see also* *Parker v. Barnhart*, 67 F. App'x. 495 (9th Cir. 2003) (finding error in an administrative law judge's failure to call a licensed psychologist, rather than another expert, as an expert witness for appropriate testimony); *Earls v. Sexton*, No. 3:09cv950, 2010 U.S. Dist. LEXIS 52980 (M.D. Pa. May 28, 2010) (allowing a nurse practitioner to testify in a negligence action concerning whether a motor vehicle accident caused psychiatric injuries); *State v. Bricker*, 321 Md. 86 (Md. Ct. App. 1990) (rejecting expert testimony from a nonpracticing psychologist who did not hold a doctorate and did not qualify for a reciprocal license under state law); *People v. McDarragh*, 175 Ill. App. 3d 284, 291 (1988) (affirming the trial court's rejection as an expert witness of a doctoral candidate who did not have the experience level required for state registration as a psychologist); *but see* *United States v. Huber*, 603 F.2d 387, 399 (2d Cir. 1979) (affirming trial court's rejection of expert testimony on defendant's mental state from a professor of economics who was also a certified psychoanalyst).

67. *See* Kirk Heilbrun, *Principles of Forensic Mental Health Assessment* 9–14 (2001) (describing differences between clinical and forensic assessments).

presentation of the findings, which may be adverse to the interests of the evaluatee.⁶⁸

Client

In a clinical assessment, the client (often termed the patient) is the person being examined and/or treated by the mental health professional, but the client in a forensic context is the person or entity that requested the evaluation (e.g., an attorney, a court, an employer, an insurer).

Relationship

Therapeutic relationships between a psychologist or psychiatrist and a patient are often characterized by empathy, support, and emotional safety. In forensic contexts, however, mental health professionals typically strive to remain detached from examinees, and the focus is on obtaining complete and accurate information that is relevant to the legal question being addressed.

Voluntariness and Autonomy

In most traditional clinical relationships, the patient enters treatment voluntarily. By contrast, in forensic contexts, examinees may be compelled to undergo an evaluation by a court, employer, or insurer or may be strongly encouraged to do so by their attorney.

Confidentiality

A patient in a traditional clinical relationship can expect that the mental health professional will not disclose confidential information obtained during treatment (unless the mental health professional is required by law to disclose certain information), but the examinee in a forensic assessment should not expect the information they provide to remain entirely confidential. The results of a forensic evaluation may be divulged to opposing counsel and in open court; in some cases, the forensic report may become part of the public record.

68. Paul S. Appelbaum, *A Theory of Ethics for Forensic Psychiatry*, 25 J. Am. Acad. Psychiatry & L. 233 (1997).

Consent

In a clinical context, the mental health professional and patient share an understanding that the goal of the therapist-patient relationship is to reduce the patient's clinical symptoms and improve their functioning. Obtaining informed consent (or the patient's agreement to participate in treatment) is a necessary part of practice. In a forensic context, the examinee's legal interests can be affected by the results of the evaluation, so it is important that the examinee understand who requested the evaluation, the purpose of the evaluation, how the results of the evaluation will be used, and who will be able to see the report. If the forensic assessment is court-ordered, the examinee is legally compelled to participate in the evaluation but should still be provided with the information noted above. But if the evaluation is not court-ordered and instead requested by the examinee's attorney or an adverse party, the examinee must be given the opportunity to decline to participate in the evaluation.

Response Style

Treatment patients are typically not motivated to conceal or distort the information they provide to a mental health professional, so mental health professionals typically assume that information provided by a patient in a treatment context is not deliberately distorted. In contrast, some examinees involved in legal proceedings may have a motivation to distort information either by exaggerating or minimizing their symptoms. It is therefore not reasonable to assume that examinees undergoing forensic assessment necessarily provide reliable information unaffected by deliberate distortion.

Data Collection and Sources of Information

Treatment relationships focus on patients' experiences, beliefs, and perceptions so that the mental health professional can better understand how patients think, feel, and experience their environment, and many mental health professionals rely almost exclusively on information provided by their patients. In forensic contexts, mental health professionals routinely consider additional sources of information (beyond self-report), typically obtaining additional information from the examinee's records (e.g., mental health, medical, criminal, school, employment) and sometimes from collateral interviews with individuals who are well acquainted with the examinee (e.g., significant others, employers, probation officers, physicians, teachers, therapists).

Pace

Traditional clinical practice typically entails a relationship between the mental health professional and patient that develops over time. In forensic contexts, evaluations are subject to deadlines associated with legal proceedings, and mental health professionals may have limited time in which to conduct an evaluation and submit a report.

Setting

Treatment typically is delivered in private, quiet settings, such as an office or a patient's hospital room. By contrast, forensic assessments of criminal defendants are often conducted in jails or prisons that offer limited privacy and poor testing conditions.

Report

In a traditional clinical context, any written report is typically brief. In a forensic context, however, mental health professionals routinely write lengthy and detailed reports that may be reviewed by attorneys, judges, and administrative bodies.

Testimony

In a treatment context, there is no expectation that mental health professionals will provide expert testimony about their patients in a deposition, hearing, or trial. In fact, fulfilling two roles in one case—such as being a treating clinician and an expert witness—may be inconsistent with professional ethical guidelines because it can result in a conflict of interest (among other problems).⁶⁹ In a forensic context, by contrast, mental health professionals should assume testimony will be required (even though testimony occurs in only a small percentage of cases).

These differences strongly suggest that treating clinicians should avoid taking on the role of forensic evaluator for a variety of reasons. Among the most important are the different purposes, procedures, limits to confidentiality, and reasons for conducting the evaluation—all of which should be conveyed in the notification of purpose or informed consent that precedes the forensic assessment

69. Thomas G. Gutheil & Paul S. Appelbaum, *Clinical Handbook of Psychiatry and the Law* 245–47 (5th ed. 2020).

itself. Another involves the importance of impartiality, which is highly valued in forensic assessment but not in clinical treatment. These various considerations are addressed in the specialty ethics guidelines for psychiatry⁷⁰ and psychology,⁷¹ which generally discourage playing both roles in the same case.

Diagnosis of Mental Disorders

Nomenclature and Typology: DSM-5 and DSM-5-TR

The standard nomenclature and diagnostic criteria for mental disorders in use in the United States are embodied in the *Diagnostic and Statistical Manual of Mental Disorders*, published by the American Psychiatric Association and now in its fifth edition with revised text (DSM-5-TR).⁷² To qualify for a DSM diagnosis, a person must meet a set of criteria characteristic of the disorder. The DSM approach has been criticized on the grounds of relevance to the evidence likely to be presented in legal proceedings, as its major goal is to provide a typology useful to clinicians and researchers that reflects the latest psychiatric understanding of mental disorders. But the DSM recognizes—in a cautionary statement in the introduction to the text—that diagnostic criteria appropriate for clinical or research purposes may not map directly onto legally relevant categories.⁷³

70. American Academy of Psychiatry and the Law, Ethics Guidelines for the Practice of Forensic Psychiatry, May 2005 [hereinafter Forensic Psychiatry Guidelines], <https://perma.cc/P88W-F7T5>.

71. American Psychological Association, Specialty Guidelines for Forensic Psychology (2011) [hereinafter APA Specialty Guidelines], <https://perma.cc/MM83-6NPZ>.

72. American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders* (5th ed. text rev. 2022) [hereinafter DSM-5-TR]. An alternative nomenclature and set of criteria used internationally can be found in the *International Classification of Diseases*, now in its eleventh edition [hereinafter ICD-11], published by the World Health Organization and available at <https://perma.cc/JSE3-WSZC>. Although the DSM-IV-TR and ICD-11 nomenclature and criteria are generally similar, there are differences that can result in diagnostic variations in particular cases, depending on which criteria are applied. The DSM-5-TR was modified in part in an attempt to make it more closely align with ICD-11.

73. The cautionary statement reads in part,

Although the DSM-5 diagnostic criteria and text are primarily designed to assist clinicians in conducting clinical assessment, case formulation, and treatment planning, DSM-5 is also used as a reference for the courts and attorneys in assessing the legal consequences of mental disorders. As a result, it is important to note that the definition of mental disorder included in DSM-5 was developed to meet the needs of clinicians, public health professionals, and research investigators rather than the technical needs of the courts and legal professionals. . . . When DSM-5 categories, criteria, and textual descriptions are employed for forensic purposes, there is a risk that diagnostic information will be misused or misunderstood. These dangers arise because of the imperfect fit between the

Since the publication of DSM-5 in 2013, an iterative revision process was established, allowing modification of criteria or text as new evidence warranting changes becomes available.⁷⁴ Hence, other than immediately after the publication of a new hard-copy edition, the canonical version of the DSM is the online version.⁷⁵

Major Diagnostic Categories

The DSM-5-TR includes more than 300 diagnoses, but the characteristics of the major categories of disorders that are likely to be relevant in legal proceedings can be summarized concisely.⁷⁶

- *Schizophrenia* is a complex psychotic⁷⁷ disorder involving delusions, hallucinations, disorganization of thought, speech, and behavior, and social withdrawal. Social and occupational functioning are markedly impaired. There may also be a deterioration in cognitive functioning, with a loss of intellectual capacity experienced over the course of the disorder, that should be appraised when schizophrenia is present. The intensity of the symptoms may fluctuate, with periodic exacerbations and often slow deterioration over time.⁷⁸
- *Bipolar disorder* (formerly called manic-depressive disorder) is a disturbance of mood marked by episodic occurrence of both mania and depression. During manic periods, persons experience elevated,

questions of ultimate concern to the law and information contained in a clinical diagnosis. In most situations, the clinical diagnosis of a DSM-5 mental disorder . . . does not imply that an individual with such a condition meets legal criteria for the presence of a mental disorder or ‘mental illness’ as defined in law, or a specified legal standard (e.g., for competence, criminal responsibility, or disability).

DSM-5-TR, *supra* note 72, at 29.

74. The guidelines for the revision process are available at <https://perma.cc/R8CP-F993>.

75. Paul S. Appelbaum, Ellen Leibenluft & Kenneth S. Kendler, *Iterative Revision of the DSM: An Interim Report from the DSM-5 Steering Committee*, 72 *Psychiatric Services* 1348 (2021), <https://doi.org/10.1176/appi.ps.202100013>.

76. These brief summaries of complex and variable conditions are meant to provide an orientation to the nature and course of major mental disorders. The current edition of the DSM itself or standard psychiatric textbooks should be consulted for more complete descriptions. Note that for a diagnosis of any disorder to be made per the DSM, the symptoms must be deemed to “cause clinically significant distress or disability in social, occupational, or other important activities.” DSM-5-TR, *supra* note 72, at 14.

77. Psychotic conditions involve some degree of detachment from reality characterized by delusional thinking and hallucinatory perceptions. *Id.* at 113.

78. *Id.* at 117–18.

expansive, or irritable mood, accompanied by such symptoms as grandiosity, racing thoughts and pressured speech, decreased sleep, and hypersexuality. The course is chronic, but intermittent, though some patients experience a downward trajectory.⁷⁹

- *Major depressive disorder* involves one or more episodes of depression, typically involving depressed mood, loss of pleasure, weight loss, insomnia, feelings of worthlessness, diminished ability to think or concentrate, and thoughts of death. Episodes are often, but not always, recurrent.⁸⁰
- *Substance use disorders* can involve a variety of substances, both licit and illicit. Recent editions of the DSM moved away from the terms *abuse* and *dependence*, and instead use the term *substance use disorder* to describe the “wide range of the disorder, from a mild form to a severe state of chronically relapsing, compulsive pattern of drug taking.”⁸¹
- *Personality disorders* are enduring patterns of inner experience and behavior that deviate markedly from the norms and expectations of the individual’s culture, are stable over time, and lead to distress or impairment.⁸² Personality disorders tend to be long-standing and difficult to treat. *Anti-social personality disorder* is often seen in criminal courts because it is marked by a pervasive pattern of disregard for and violation of the rights of others.⁸³
- *Neurocognitive disorders*, which include the conditions previously referred to as *dementia*, are marked by impairment of cognitive abilities, including memory, language, motor functions, recognition of objects, and executive functioning.⁸⁴ The most common form of progressive neurocognitive disorder is Alzheimer’s disease, the incidence of which increases with age and the cause of which remains unclear, although in many cases, genetics seems to play a role.⁸⁵ Other causes of neurocognitive disorders include multiple small strokes (*multi-infarct dementia*), trauma, and infection with certain virus-like agents.
- *Intellectual disability* (formerly referred to as mental retardation) has onset during the developmental period and includes both intellectual and adaptive functioning deficits in conceptual, social, and practical domains.⁸⁶

79. *Id.* at 139–59.

80. *Id.* at 183–92.

81. *Id.* at 543.

82. *Id.* at 733.

83. *Id.* at 748–52.

84. *Id.* at 679–80.

85. Rebecca Sims, Matthew Hill & Julie Williams, *The Multiplex Model of the Genetics of Alzheimer’s Disease*, 23 *Nat’l Neuroscience* 311 (2020), <https://doi.org/10.1038/s41593-020-0559-5>.

86. DSM-5-TR, *supra* note 72, at 37–38.

- *Post-traumatic stress disorder* (PTSD) can develop following exposure to actual or threatened death, serious injury, or sexual violence. Symptoms may include avoidance of experiences that are associated with the trauma, intrusive and disturbing thoughts and feelings about it, negative changes in thinking and mood, and more frequently experienced arousal (e.g., fight-or-flight reactions). It may be seen in cases in which youth have experienced abuse and neglect, for example, or in cases involving intimate partner violence or sexual assault, combat, or accidental injury.

Additional disorders that may have special legal relevance include dissociative disorders (such as dissociative identity disorder, formerly multiple personality disorder), impulse-control disorders (such as kleptomania and pyromania), sexual-behavior disorders (especially the paraphilias, such as pedophilia), and delirium.⁸⁷

Response Style

Malingering and Exaggeration

Because the diagnosis of mental disorders rests heavily on the elicitation of symptoms from the person being evaluated and observations of the person's behavior, the possibility of malingering—the deliberate simulation of symptoms of mental disorder—must always be considered.⁸⁸ Most commonly, the likelihood of malingering is assessed as part of a clinical evaluation. The pattern of symptoms

87. *Rebrook v. Astrue*, No. 5:07CV39, 2008 WL 822104 (N.D. W. Va. Mar. 26, 2008) (anxiety disorder); *United States v. Holsey*, 995 F.2d 960 (10th Cir. 1993) (dissociative disorder); *Coe v. Bell*, 89 F. Supp. 2d 922 (M.D. Tenn. 2000) (dissociative identity disorder); *United States v. Miller*, 146 F.3d 1281 (11th Cir. 1998) (impulse-control disorder); *United States v. McBroom*, 991 F. Supp. 445 (D.N.J. 1998) (person receiving treatment for bipolar disorder and impulse-control disorder sentenced for possession of child pornography); *United States v. Silleg*, 311 F.3d 557 (2d Cir. 2002) (pedophilia determination in a child-pornography case); *Fields v. Lyng*, 705 F. Supp. 1134 (D. Md. 1988) (kleptomania); *United States v. Warr*, 530 F.3d 1152 (9th Cir. 2008) (sentencing of an arsonist diagnosed with pyromania upheld); *Kansas v. Hendricks*, 521 U.S. 346 (1997) (upholding commitment of man unable to control pedophilic impulses); *United States v. Gigante*, 996 F. Supp. 194 (E.D.N.Y. 1998) (dementia); *Johnson v. City of Cincinnati*, 39 F. Supp. 2d 1013 (S.D. Ohio 1999) (estate of man who died from police restraint during a seizure sued the city under 28 U.S.C. § 1983); *Bertl v. City of Westland*, No. 04-CV-75001, 2007 WL 3333011 (E.D. Mich. Nov. 9, 2007) (finding that delirium tremens is an objectively serious medical need); *Atkins v. Virginia*, 536 U.S. 304 (2002) (banning the execution of the mentally retarded as a violation of the Eighth Amendment); *In re Hearn*, 418 F.3d 444 (5th Cir. 2005); *Hamilton v. Southwestern Bell Tel. Co.*, 136 F.3d 1047, 1050 (5th Cir. 1998) (recognizing PTSD as a mental impairment for the purposes of the Americans with Disabilities Act).

88. Malingering is defined as “the intentional production of false or grossly exaggerated physical or psychological symptoms, motivated by external incentives such as avoiding military duty, avoiding work, obtaining financial compensation, evading criminal prosecution, or obtaining

reported by the person is compared with known syndromes, and the consistency of the person's behaviors is observed. Contrary to common belief, mental disorders are not easy to fake, especially when the deception must be sustained over a period of time.⁸⁹

When deception is suspected, efforts to confirm it should begin during the clinical examination, as the person is offered the opportunity to endorse symptoms that are unlikely to occur naturally (e.g., "Do you ever feel as though the cars on the street are talking about you?") or that do not fit the condition from which the patient is claiming to suffer.⁹⁰ Psychological testing can be helpful in detecting deception; the Minnesota Multiphasic Personality Inventory-3 (MMPI-3), for example, has scales that correlate with persons who are both "faking bad" (i.e., fabricating symptoms) and "faking good" (i.e., hiding symptoms that actually exist).⁹¹ Other instruments specifically for the assessment of malingering also have been developed, with varying degrees of validation.⁹² Information from records of previous psychiatric or psychological evaluations can be helpful in determining the congruence of a person's current symptoms with past reports and behaviors. In addition, given the difficulty in maintaining a consistent pattern of deception over a sustained period, data provided by collateral sources (e.g., family members, roommates, prisoners in adjoining cells, correctional officers, nurses and other hospital staff) who have observed the person informally can be crucial in distinguishing real from malingered disorders.⁹³

The difficulty of simulating a mental disorder does not imply that it is impossible to do. Indeed, a skilled and determined person can sometimes fool even an experienced evaluator. Thus, the only honest response that a clinician can give in almost every circumstance to a question about the possibility of malingering is that it is always possible, but is more or less likely in this particular case, given the characteristics of the person being evaluated.⁹⁴

drugs," DSM-5-TR, *supra* note 72, at 835; see Phillip J. Resnick, *Malingering*, in *Principles and Practice of Forensic Psychiatry* 543 (Richard Rosner ed., 2003).

89. See Resnick, *supra* note 88, at 544.

90. Gutheil & Appelbaum, *supra* note 69, at 258–59.

91. See Claudia K. Reeves, Tiffany A. Brown & Martin Sellbom, *An Examination of the MMPI-3 Validity Scales in Detecting Overreporting of Psychological Problems*, 34 *Psych. Assessment* 517 (2022), <https://doi.org/10.1037/pas0001112> (noting MMPI-3 validity scales can effectively differentiate malingering from genuine mental disorder); Megan R. Whitman, Jessica L. Tylicki & Yossef S. Ben-Porath, *Utility of the MMPI-3 Validity Scales for Detecting Overreporting and Underreporting and Their Effects on Substantive Scale Validity: A Simulation Study*, 33 *Psych. Assessment* 411 (2021), <https://doi.org/10.1037/pas0000988> (discussing MMPI-3's ability to detect overreporting and underreporting of psychological symptoms).

92. See generally Rogers & Bender, *supra* note 20 (discussing instruments and approaches for detecting malingering).

93. Gutheil & Appelbaum, *supra* note 69, at 259.

94. Resnick, *supra* note 88.

Other Response Styles

The literature on response styles has expanded quite considerably in recent years, and researchers have now identified several different types.⁹⁵ For example, the response style of defensiveness is characterized as the polar opposite of malingering—that is, the denial or gross minimization of psychological and physical symptoms with the motivation of obtaining an unwarranted, yet desired, external incentive.⁹⁶ Social desirability is a response style in which individuals present themselves in the most favorable manner relative to social norms.⁹⁷ An uncooperative response style describes individuals who deliberately withhold information during the evaluation process.⁹⁸ An irrelevant response style describes an individual who does not sufficiently engage in the assessment process,⁹⁹ which can limit the evaluator's ability to draw valid conclusions. Acquiescent responding, which is sometimes called yea-saying, is when an individual responds in the affirmative to most or all questions/items, while disacquiescent responding, or nay-saying, is the opposite response style.¹⁰⁰ Hybrid responding describes an individual who uses more than one response style during an evaluation.¹⁰¹ All of these response styles are distinguishable from reliable responding, which is when individuals respond genuinely to evaluation items and questions.

Functional Impairment Due to Mental Disorders

Impact of Mental Disorders on Functional Capacities

Mental disorders can affect functional capacities in various ways. Among these, attention and concentration may be impaired by the preoccupations that appear in anxiety and depressive disorders, or the grosser distractions (e.g., auditory hallucinations) of psychotic disorders.¹⁰² Perception is often distorted in psychotic conditions, as manifested by hallucinations of the auditory, visual, tactile, or other

95. See Rogers & Bender, *supra* note 20 (providing comprehensive overview of malingering and other response styles).

96. See Richard Rogers, *An Introduction to Response Styles*, in Rogers & Bender, *supra* note 20 (discussing various response styles besides malingering).

97. See *id.* at 7.

98. See *id.* at 10.

99. See *id.* at 8.

100. See *id.*

101. See *id.*

102. James G. Scott, *Attention/Concentration: The Distractible Patient*, in *The Little Black Book of Neuropsychology* (Mike R. Schoenberg & James G. Scott eds., 2011).

sensory systems.¹⁰³ Cognition, encompassing both the process and content of thought, is also often affected: Thought processes can be impeded by the slowing of thought in depression, its acceleration in mania, or the scrambling of thought among persons with schizophrenia or other psychotic disorders. Thought content may be altered by the odd reasoning to which persons with delusions appear to be prone.¹⁰⁴ Motivation to act, even in one's self-interest, is often globally reduced in states of intense depression and in schizophrenia.¹⁰⁵ Judgment and insight may be altered under the pressure of delusions.¹⁰⁶ Control of behavior can be weakened by the impulsivity seen in mania and psychosis, the drives of the impulse disorders, and the use of disinhibiting substances, especially alcohol.¹⁰⁷ Any of these impairments could affect a person's relevant decisional and performative capacities.

This necessarily incomplete list of the ways that mental disorders can affect functional capacities illustrates the vulnerability of almost every aspect of mental functioning to perturbation. And although it is common to divide mental functions into categories such as these for heuristic purposes, most neuroscientists recognize that the brain operates as a unified entity.¹⁰⁸ Thus, it is rare that impairments are limited to a single area of functioning. Impaired concentration, for example, inherently affects cognitive abilities, which in turn may alter judgment and therefore the person's choice of behaviors. Although focal deficits may occur, for example the anxiety associated with exposure to a phobic stimulus such as a spider, more severe disorders will have a broader impact on a person's functional capacities as a whole.¹⁰⁹

Assessment of Functional Legal Capacities

Determining the nature and extent of past, present, or future functional impairment, therefore, is usually the most critical aspect of a forensic mental health evaluation and subsequent presentation of mental health evidence.

103. Andre Aleman & Frank Laroi, *Hallucinations: The Science of Idiosyncratic Perception* (2008).

104. Joel Yager & Michael J. Gitlin, *Clinical Manifestations of Psychiatric Disorders*, in 10 *Comprehensive Textbook of Psychiatry* 973–80 (Benjamin J. Sadock, Virginia A. Sadock & Pedro Ruiz eds., 2017).

105. *Id.* at 984.

106. Phillipa A. Garety, *Insight and Delusions*, in *Insight and Psychosis* 66, 66–77 (Xavier F. Amador & Anthony S. David eds., 1998).

107. Eric Hollander & J. Rosen, *Impulsivity*, 14 *J. Psychopharmacology* S39 (2000).

108. William R. Uttal, *The New Phrenology: The Limits of Localizing Cognitive Processes in the Brain* (2003).

109. The pervasive impact of schizophrenia on all aspects of personality and functioning is the most extreme example. See Ryan Lawrence et al., *Schizophrenia Spectrum and Other Psychotic Disorders*, in Joseph M. Roberts et al., *The Utility of the Trauma Symptom Inventory as a Primary and Secondary Assessment Instrument for Forensic Practice in Legal Settings* 257–78 (2022).

Clinical Examination

As in establishing a diagnosis, the core of the assessment of functional impairment remains the clinical examination.¹¹⁰ A diagnostic assessment may be integral to the functional assessment process, suggesting to the examiner areas of possible impairment to be explored in greater depth (e.g., attentional and concentration abilities in an anxiety disorder; impairments in motivation in a depressive disorder). Beginning in the 1970s, however, there was growing recognition among the mental health professions that merely establishing a diagnosis is not sufficient for drawing a conclusion about a legally relevant capacity, because a broad range of functional impairments can be associated with almost any mental disorder.¹¹¹

Thus, in addition to a diagnostic assessment, an adequate examination will explore the person's perspective on the alleged functional impairment and will probe for symptoms associated with such impairment. The process involves more than simply taking the person's word for the issue in question, for example, that someone was not able to comprehend the details of the contract to which they are a party or that they remain incapable of the careful calculations required in their job. Assessors compare the claimed impairments with the person's overall history and other areas of function, looking for congruence or incongruence. For example, the assertion by a plaintiff that because of being harassed on the job, the plaintiff has been unable to concentrate sufficiently to work will be more or less plausible depending on the consistency and extent of the symptoms and the degree to which the impairment may generalize to other areas of the person's life. Degrees of claimed impairment that are out of scale with the extent of symptoms or the person's functional history are inherently suspect.¹¹²

In addition to questioning the evaluatee directly, the use of collateral information can be essential to a valid assessment, particularly when the person has an incentive to malingering, which will often be the case in legal proceedings.¹¹³ Family members, coworkers, and others who have had an opportunity to observe the person can provide invaluable information about the nature and extent of impairments, although one must always be alert to the possibility that informants will be motivated to assist the person by distorting or exaggerating their accounts. Records of performance, such as educational test results and work evaluations, especially if generated prior to the filing of the legal claim, may shed somewhat more objective light on the person's capacities.¹¹⁴ To the extent that impairments may

110. See section titled "Diagnosis of Mental Disorders" above.

111. Michael Kindred, *Guardianship and Limitations upon Capacity*, in *The Mentally Retarded Citizen and the Law* 62 (The President's Comm. on Mental Retardation 1976); Lab'y of Cmty. Psychiatry, Harvard Med. Sch., *Competency to Stand Trial and Mental Illness* (1973).

112. Rogers & Bender, *supra* note 20.

113. Gutheil & Appelbaum, *supra* note 69.

114. Melton et al., *supra* note 1, at 53–55.

be rooted in disruptions of brain functions per se, neuropsychological testing can also be helpful in documenting their nature and extent. Increasingly, however, an unstructured clinical evaluation on its own, even when supplemented by collateral information, is not necessarily the most accurate tool for determining functional capacity.¹¹⁵

Structured Assessment Techniques

As with determination of diagnosis, the evaluation of the limitations of function due to mental disorders increasingly involves the use of structured assessment techniques.¹¹⁶ Most commonly, these are standardized interviews or data-gathering protocols (e.g., based on a person's psychiatric record) designed to ensure that all relevant information is obtained. In addition, where research has established the validity of the instruments by demonstrating a correlation between the results and actual impairments, these techniques may allow a quantitative estimate of the extent of actual functional deficiencies to be made. A compendium of assessment instruments includes structured evaluations that address criminal defendants' competence to stand trial, waiver of rights to silence and legal counsel, criminal responsibility, one's parenting capacity, competence to manage one's affairs (i.e., need for a guardian or conservator), and competence to consent to medical treatment and research.¹¹⁷ Given that this area is a rapidly developing focus of research, instruments to address other legally relevant functional capacities and states—propensity to commit violent or sexual offenses comes quickly to mind¹¹⁸—are continuously being tested and developed.

115. Grisso, *supra* note 3. Surprisingly few studies exist of the reliability of clinical forensic evaluations. An early U.S. study of actual assessments showed good interrater reliability of evaluations of competence to stand trial, although many of the reports were deficient in other ways. Jennifer L. Skeem et al., *Logic and Reliability of Evaluations of Competence to Stand Trial*, 22 L. Hum. Behav. 519 (1998), <https://doi.org/10.1023/a:1025787429972>. An Australian study found only fair to moderate reliability across assessments of competence to stand trial, but moderate to good reliability of criminal-responsibility evaluations. Matthew Large et al., *Reliability of Psychiatric Evidence in Serious Criminal Matters: Fitness to Stand Trial and the Defence of Mental Illness*, 43 Austl. N.Z. J. Psychiatry 446 (2009), <https://doi.org/10.1080/00048670902817745>. A systematic review and meta-analysis found relatively poor reliability in both competence-to-stand trial and insanity evaluations. See Lucy A. Guarnera & Daniel C. Murrie, *Field Reliability of Competency and Sanity Opinions: A Systematic Review and Meta-Analysis*, 29 Psych. Assessment 795 (2017), <https://doi.org/10.1037/pas0000388>.

116. See Grisso, *supra* note 3.

117. *Id.*

118. See, e.g., Kevin S. Douglas et al., HCR-20 V3 Assessing Risk for Violence User Guide, Version 3 (2013); Grant T. Harris et al., *Violent Offenders: Appraising and Managing Risk* (2015); John Monahan et al., COVR—Classification of Violence Risk, Professional Manual (2005).

Although most assessment techniques rely on information gathered from the person being evaluated or from existing records, some approaches involve direct testing of the person's capacity to perform particular tasks. Examples include computerized assessment of driving capacities,¹¹⁹ observation of tasks involving the handling and management of money¹²⁰ and of parenting skills,¹²¹ and direct measurement of such capacities as understanding and reasoning about medical information when a person's competence to decide about medical treatment is at issue.¹²² In general, these approaches reduce the degree of inference required in drawing conclusions about a person's functioning because the person is observed performing something close to the precise tasks in question. Of course, such techniques may not be relevant when the legal issue relates to the impact of mental disorder on functional abilities at some time in the past or future, especially if the person's mental state at present may be different from what it was or will be. Nonetheless, these can be useful approaches to evaluation in appropriate legal contexts.

The advantages that attend the use of structured assessment instruments include the thoroughness of the evaluation, and in many cases, a research base exists from which conclusions can be drawn about the degree of an assessed person's functional impairment.¹²³ In some jurisdictions, the use of structured assessments is even required for particular purposes (e.g., evaluation of sex offenders).¹²⁴ Still, the use of structured assessments performed for the purpose of being introduced in legal proceedings is variable and far from universal.¹²⁵ Grisso, a leading scholar in this area, suggests three reasons why: (1) it is easier and may be more lucrative (e.g., where a fixed rate is being paid per evaluation) for an examiner to avoid the frequently time-consuming use of a structured instrument; (2) many cases involve persons whose functional impairments—or lack of impairment—are obvious, and use of a structured assessment instrument would be overkill; and (3) perhaps paradoxically, the use of an assessment tool

119. Maria T. Schultheis et al., *The Neurocognitive Driving Test: Applying Technology to the Assessment of Driving Ability Following Brain Injury*, 48 *Rehabilitation Psych.* 275 (2003), <https://doi.org/10.1037/0090-5550.48.4.275>.

120. Dan Marson et al., *Assessing Financial Capacity in Patients with Alzheimer's Disease: A Conceptual Model and Prototype Instrument*, 57 *Archives Neurology* 877 (2000), <https://doi.org/10.1001/archneur.57.6.877>.

121. Marc J. Ackerman & Kathleen Schoendorf, *Ackerman-Schoendorf Scales for Parent Evaluation of Custody Manual* (1992).

122. Thomas Grisso & Paul S. Appelbaum, *MacArthur Competence Assessment Tool for Treatment (MacCAT-T)* (1998).

123. Grisso, *supra* note 3, at 45–47.

124. See, e.g., Va. Code Ann. § 37.2-903-C: “Each month the Director shall review the database and identify all such prisoners who are scheduled for release from prison within 10 months from the date of such review who receive a score of five or more on the Static-99 or a like score on a comparable, scientifically validated instrument designated by the Commissioner. . . .”

125. Grisso, *supra* note 3, at 481.

makes experts more vulnerable to attack on cross-examination.¹²⁶ That many expert witnesses lack knowledge of the existence of these instruments and that a sense that their use denigrates the evaluator's expertise should be added to this list.

The vulnerability of testimony based on assessment instruments to cross-examination is worth special emphasis. Opinions offered on the basis of "clinical experience," which appears to be the norm, are difficult to challenge when expert witnesses in fact have appropriate training and a good deal of experience with the condition in question.¹²⁷ On the other hand, assessment instruments can be subjected to scrutiny with regard to the empirical database that supports their use, including their reliability and validity, their acceptance by the relevant professional community, and their probative value in a particular case. There may also be questions regarding the examiner's training and experience with the instrument and whether it was administered in the manner intended by its developers. All of these are legitimate questions, of course, and an argument can be made that the introduction of data from assessment instruments into evidence should be held to a more rigorous standard, because factfinders may give such data greater credence than unassisted clinical judgment.¹²⁸ But the undoubted consequence is that the arguably more reliable and perhaps more valid data from empirically derived assessment techniques are less likely to be introduced in evidence than evaluators' subjective judgments of unknown validity.¹²⁹

Predictive Assessments

As noted above,¹³⁰ predictive assessments are the most challenging evaluations performed by mental health professionals.¹³¹ The most common predictive tasks involve the prediction of violence and of future functional impairment and

126. *Id.* at 481–82.

127. See Jack B. Weinstein, 4 Weinstein's Federal Evidence § 702:02 n.1 (2d ed. 2008) (on the liberal admissibility of expert testimony under Federal Rule of Evidence 702); § 702.02[4] nn.25–27 (on the trial judge's broad discretion to admit or exclude expert testimony and determine its helpfulness and relevancy, and on the application of the "abuse of discretion" standard of review to determinations of whether a witness qualifies as an expert); § 702.04[1][c] (on the typical "academic credentials plus experience" combination). *Bryan v. City of Chicago*, 200 F.3d 1092 (7th Cir. 2000) (an expert may qualify based on academic expertise and practical experience).

128. Christopher Slobogin, *Experts, Mental States, and Acts*, 38 Seton Hall L. Rev. 1009 (2008).

129. Grisso, *supra* note 3, at 482.

130. See section titled "Overview of Mental Health Evidence" above.

131. Yogi Berra, New York Yankees' Hall of Fame catcher and philosopher of everyday life, is purported to have said, "It's tough to make predictions, especially about the future." See <https://perma.cc/5FLR-WJ6T>. For a discussion of the origin of the Berra phrase, see Henry T. Greely and Anthony D. Wagner, *Reference Guide on Neuroscience*, in 3 Reference Manual on Scientific Evidence (2011).

responses to treatment. A variation on the question of prediction involves appraising the domains of risk, need, and responsivity. These professional activities are discussed in the sections that follow.

Prediction of Violence

The risk that a person may commit a violent act at some point in the future may come into play in the criminal process when courts are making determinations of suitability for diversion, bail, sentencing, probation, and parole, and in the civil process in hearings for civil commitment to psychiatric facilities and sex-offender treatment programs and when considering the imposition of liability on clinicians and facilities for failing to protect victims of patients' violence.¹³² Not all persons for whom such assessments must be made will have mental disorders, but psychiatrists and psychologists are seen by the courts as having expertise in this area and hence are almost invariably called upon for these evaluations.¹³³

Persons with serious mental disorders, such as schizophrenia or bipolar disorder, are often considered by the general public to be at high risk for violence.¹³⁴ However, data on the relationship between serious mental disorders and violence are variable (schizophrenia is the disorder most frequently studied). Although most studies suggest a moderately elevated risk, the proportion of violence accounted for by individuals with serious mental disorders is small, probably 3% to 5%, based on the best available U.S. estimates.¹³⁵ Data also suggest that the stereotype of individuals with serious mental disorders who assault strangers in

132. See, e.g., *Kansas v. Hendricks*, 521 U.S. 346 (1997); *White v. Johnson*, 153 F.3d 197 (5th Cir. 1998). A unanimous U.S. Supreme Court in *Chambers v. United States*, 129 S. Ct. 687, 691–93 (2009), pointed to the importance of considering empirical data when identifying circumstances associated with increased risk of violence.

Violence can be defined in different ways. One definition frequently used in criminal law involves the commission of an offense against persons. Another definition employed by researchers is some variation on *serious acts of violence* (battery that resulted in physical injury; sexual assaults; assaultive acts that involved the use of a weapon; or threats made with a weapon in hand) and *other aggressive acts* (battery not resulting in physical injury or threats without a weapon in hand), a distinction used by the MacArthur Violence Risk Assessment Study Group. See <https://perma.cc/45H2-3YK9>.

133. See Joanmarie Ilaria Davoli, *Psychiatric Evidence on Trial*, 56 S.M.U. L. Rev. 2191 (2003).

134. See Bernice Pescosolido et al., *The Public's View of the Competence, Dangerousness, and Need for Legal Coercion Among Persons with Mental Illness*, 89 Am. J. Pub. Health 1339 (1999), <https://doi.org/10.2105/AJPH.89.9.1339>; see also P.W. Corrigan et al., *Implications of Educating the Public on Mental Illness, Violence, and Stigma*, 55 Psychiatric Servs. 577 (2004), <https://doi.org/10.1176/appi/ps.55.5.577>.

135. Eric Elbogen & Sally Johnson, *The Intricate Link Between Violence and Mental Disorder*, 66 Arch. Gen. Psychiatry 152 (2009), <https://doi.org/10.1001/archgenpsychiatry.2008.537>; Paul S. Appelbaum, *Violence and Mental Disorders: Data and Public Policy* (editorial), 163 Am. J. Psychiatry 1319 (2006).

public places is inaccurate: Most violence by persons with serious mental disorders is directed at family members and friends and occurs in the living quarters of the perpetrator or the victim.¹³⁶ Much higher rates of violence are associated with substance use,¹³⁷ especially alcohol use, and with psychopathy, a subcategory of antisocial personality disorder.¹³⁸ Indeed, most of the strongest predictors of violence are common to both persons with serious mental disorders and those without, suggesting that the impact of the disorders per se is slight.¹³⁹

Approaches to Prediction of Violence

Clinical evaluation of violence risk ordinarily focuses on those variables that have been shown in empirical research to have the strongest relationship to future violence,¹⁴⁰ whether the information is gleaned directly from the person or derived from collateral informants or from a review of relevant records. These variables include a history of previous violence, age (violence risk peaks in the late teens and early twenties, declines slowly through the twenties and thirties, and drops off precipitously after age forty), male sex, lower socioeconomic status, employment instability, substance misuse, psychopathic personality traits, and childhood victimization.¹⁴¹ The evaluation process is complicated by the fact that literally scores of variables show some significant relationship with future violence—but this relationship is not strong enough or sufficiently independent of other variables to contribute to the aggregation of an accurate prediction of violence.¹⁴² However, beginning with the variables noted above, the evaluator estimates the baseline risk of violence for the person and then adjusts that value by taking into account foreseeable changes in relevant areas during the outcome period. When previous violence has occurred, the risk estimate is adjusted to include those specific variables that have been associated with violence by this person in the past (e.g., being left by a girlfriend), including whether they are present at the time of evaluation or likely to recur in the future.¹⁴³

136. Henry J. Steadman et al., *Violence by People Discharged from Acute Psychiatric Inpatient Facilities and by Others in the Same Neighborhoods*, 55 Arch. Gen. Psychiatry 393 (1998), <https://doi.org/10.1001/archps.55.5.393>.

137. Elbogen & Johnson, *supra* note 135.

138. John Monahan et al., *Rethinking Risk Assessment: The MacArthur Study of Mental Disorder and Violence* (2001).

139. *Id.* at 37–90; Jennifer L. Skeem et al., *Offenders with Mental Illness Have Criminogenic Needs, Too: Toward Recidivism Reduction*, 38 L. Hum. Behav. 212 (2014), <https://doi.org/10.1037/lhb0000054>.

140. Gutheil & Appelbaum, *supra* note 69, at 59.

141. *Id.*

142. Monahan et al., *supra* note 138, at 163–68.

143. Gutheil & Appelbaum, *supra* note 69.

A growing number of structured assessment instruments specific to the prediction of future violence risk have been developed in the past several decades. Among the best known of these are the Historical Clinical Risk Management–20, Version 3 (HCR–20–V3),¹⁴⁴ the Violence Risk Assessment Guide (VRAG),¹⁴⁵ and the computerized Classification of Violence Risk (COVR).¹⁴⁶ A set of instruments also exists for the prediction of future sex offending.¹⁴⁷ Violence risk-assessment instruments have been developed in one of three ways: by assembling known predictors from the research literature and applying them using professional judgment (e.g., HCR–20–V3), by assembling known predictors from the research literature and applying them using empirically guided rules using data from large subject populations (e.g., the Violence Risk Scale¹⁴⁸), or drawing and applying known predictors using statistical analysis of research data from large subject populations (e.g., VRAG and COVR). After a measure is developed, researchers validate it using populations similar to the ones with which it is anticipated they will be used. The more sophisticated measures yield estimates of the degree of risk, using either categories (e.g., high, medium, and low risk) or probabilities (e.g., 10%, 25%, or 40% likely) rather than dichotomous predictions that violence will or will not occur. Validation research must demonstrate a relationship between the estimated degree of risk and the probability of future violence for the instrument to be considered empirically supported.¹⁴⁹

The literature on prediction is marked by strong and unresolved differences of opinion over the best basis for the ultimate risk estimate. Those who support exclusive reliance on the quantitative predictions generated by structured assessment instruments, which is often referred to as actuarial prediction, argue that any attempts to modify the resulting risk estimates would necessarily reduce accuracy.¹⁵⁰ Proponents of using professional judgment to contribute to risk assessment note that exclusive reliance on instrumentation is unwise because of the inevitable questions about how applicable the group data, on which an instrument is based, are to the person being evaluated. They also argue that a fixed set of questions can never capture all the variables that may be relevant in a particular

144. Douglas et al., *supra* note 118.

145. Harris et al., *supra* note 118.

146. Monahan et al., *supra* note 118.

147. Handbook of Violence Risk Assessment (Kevin S. Douglas & Randy K. Otto eds., 2021).

148. Stephen C. Wong & A.E. Gordon, *The Validity and Reliability of the Violence Risk Scale: A Treatment-Friendly Violence Risk Assessment Tool*, 12 Psych. Pub. Pol. & L. 279 (2006), <https://doi.org/10.1037/1076-8971.12.3.279>.

149. Harris et al., *supra* note 118; John Monahan et al., *An Actuarial Model of Violence Risk Assessment for Persons with Mental Disorders*, 56 Psychiatric Services 810 (2005), <https://doi.org/10.1176/appi.ps.56.7.810>.

150. N. Zoe Hilton et al., *Sixty-Six Years of Research on the Clinical Versus Actuarial Prediction of Violence*, 34 Counseling Psych. 400 (2006), <https://doi.org/10.1177/0011000005285877>.

situation and that evaluatees are sometimes uncooperative with a structured process.¹⁵¹ Compromises include using an actuarial measure but allowing clinical judgment in interpreting the results in light of additional considerations, or using a measure that has been designed to structure the evaluation around relevant risk factors but that requires the user to reach a final structured professional judgment (SPJ) on the basis of the totality of the information. There is a robust literature on actuarial and SPJ approaches to risk assessment, with the most recently aggregated evidence suggesting that they are comparable in predictive accuracy regarding future violence.¹⁵²

Limitations of Violence Prediction

There is voluminous research literature on violence prediction. Studies of predictions by psychiatrists and psychologists in the 1960s and 1970s showed poor accuracy in judging whether persons with mental disorders and sex offenders would be likely to be violent at some point after release.¹⁵³ Indeed, the most frequently cited conclusion was Monahan's statement that when mental health professionals predicted that a person would be violent, they were twice as likely to be wrong as right.¹⁵⁴ The cumulative impact of these findings stimulated a great deal of research to identify variables that predict violence and incorporate them into both clinical predictions and the structured assessment instruments described above.¹⁵⁵

At this point, it is possible to identify several items of consensus from the research literature. Violence is not a unitary phenomenon; that is, it occurs for different reasons, related both to the motivations of the perpetrator and to the environmental context.¹⁵⁶ A bar room brawl has different roots than a mugging; the precipitants of spouse abuse bear little similarity to the motivations underlying a killing that has been premeditated as an act of revenge. Thus, no single variable or set of variables can be relied upon in all cases to appraise violence risk. Long-term prediction of violence is less accurate, both because some kinds of violence are rare¹⁵⁷ and because it is difficult for clinicians to anticipate changes in the person and the environment over time and their effects on the person's

151. Kirk S. Heilbrun et al., *Approaches to Violence Risk Assessment: Overview, Critical Analysis, and Future Directions*, in Douglas & Otto, *supra* note 147.

152. *Id.*

153. John Monahan, *The Clinical Prediction of Violent Behavior* (1981).

154. *Id.* at 60.

155. Heilbrun et al., *supra* note 151.

156. Paul S. Appelbaum, *Preface*, in *Clinical Assessment of Dangerousness: Empirical Contributions ix–xiv* (Georges-Franck Pinard & Linda Pagani eds., 2001).

157. Paul E. Meehl, *Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence* (1954).

behavior.¹⁵⁸ On the other hand, shorter-term prediction (i.e., days to weeks) has more potential to be accurate. It is worth noting that even when the leading actuarial instruments are used to make dichotomous judgments of future violence—that is, a cutoff is set to simulate the clinical prediction process—their rates of accuracy are similar.¹⁵⁹ Mental health professionals, therefore, have been encouraged to move away from attempting to make dichotomous judgments of dangerousness and toward predictions couched in terms of the risk of future violence,¹⁶⁰ and the field has moved strongly in this direction. Even here, though, high degrees of accuracy have not been attained—and may be unattainable. The state of the art probably allows well-trained clinicians, especially if they are using structured assessment instruments, to assign persons into high-, medium-, and low-risk groups with reasonable accuracy. At present, the hope of designating risk categories with greater precision for most categories of persons with mental disorders is likely illusory.¹⁶¹ When quantitative data are available, however, precision in communication of risk would undoubtedly be enhanced if the data were utilized and if assessors specified their definitions of the categories being employed.¹⁶²

The studies on the accuracy of prediction, whether clinical or actuarial, have typically involved the direct evaluation of the person about whom the prediction was being made. Opinions about the risk of future violence by persons whom the evaluator has not examined are more problematic. Although risk can

158. Jennifer L. Skeem et al., *Building Mental Health Professionals' Decisional Models into Tests of Predictive Validity: The Accuracy of Contextualized Predictions of Violence*, 24 L. & Hum. Behav. 607 (2000), <https://doi.org/10.1023/a:1005513818748>.

159. Heilbrun et al., *supra* note 151.

160. Henry J. Steadman et al., *From Dangerousness to Risk Assessment: Implications for Appropriate Research Strategies*, in *Mental Disorder and Crime* (Sheilagh Hodgins ed., 1993).

161. One method for determining the accuracy of risk-assessment measures such as those discussed in this reference guide is receiver operating characteristics (ROC), a statistical means of distinguishing between those who have the outcome (e.g., violence) and those who do not. The area under the curve (AUC) value summarizes ROC-calculated discriminatory capacity, and most good risk-assessment measures have AUC values between .70 and .80. An AUC value of .80 means that the risk-assessment measure in 80% of cases assigns a higher risk to individuals who eventually show violence than to individuals who do not. See Douglas & Otto, *supra* note 147.

162. See Kelly M. Babchishin & R. Karl Hanson, *Improving Our Talk: Moving Beyond the "Low," "Moderate," and "High" Typology of Risk Communication*, 16 Crime Scene 11 (2009). Suggestions for improving the clarity of risk communications include distinguishing between the likelihood of future violence and the anticipated severity of the offense, specifying the period for which the prediction is being made (e.g., "over the next six months"), indicating the comparison population for the estimate (e.g., "risk is high compared with the general population" or "risk is high compared with the population of persons with similar histories of violence"), and providing both absolute and relative risks when quantitative data are available (e.g., "risk of future violence over the next year is between 8 and 12%, which is between 4 and 6 times greater than would be expected for the general population").

be calculated from records only using some measures such as the STATIC-99¹⁶³ or the Psychopathy Checklist—Revised,¹⁶⁴ personal contact with an individual as part of a professional risk assessment is important for several reasons. First, it allows the evaluator to obtain informed consent or provide notification of purpose. Second, it permits the evaluator to “individualize” the appraisal by considering the individual’s appearance, responses to detailed questions, severity of symptoms, life circumstances, and other considerations, which is important in legal settings but recognized within the field as challenging (having been termed the group-to-individual problem).¹⁶⁵ Third, it facilitates the process by which the evaluator considers various possibilities and accepts or rejects them depending upon the observed clinical data. For these reasons, personal contact can reduce reliance on inaccurate information and make the overall conclusion more accurate and credible.

Opinions about the future dangerousness of individuals who have not been personally evaluated have been introduced, for example, in death penalty cases in which the prosecution sought to prove that further violence was likely, but the defense denied the prosecution expert direct access to the defendant.¹⁶⁶ If such evidence is to be introduced, at a minimum, one would expect the assessor to note the limitations on the assessor’s knowledge of the evaluatee and on the certainty with which conclusions can be reached.

Predictions of Future Functional Impairment

Cases involving claims of emotional harms, along with disability and workers’ compensation claims, often require estimates of the plaintiff’s future functional impairment so that damages can be determined accordingly.¹⁶⁷ Techniques for the assessment of function were described above.¹⁶⁸ However, these cases call for something more: predictions of the degree of change in functional impairment

163. See <https://perma.cc/7E6U-2V7S>.

164. Robert D. Hare, *The Psychopathy Checklist—Revised* (2d ed. 2003).

165. David L. Faigman, John Monahan & Christopher Slobogin, *Group to Individual (G2i) Inference in Scientific Expert Testimony*, 81 U. Chi. L. Rev. 417 (2014).

166. See *Barefoot v. Estelle*, 463 U.S. 880 (1983); Ron Rosenbaum, *Travels with Doctor Death*, Vanity Fair, May 1990, at 141.

167. 20 C.F.R. § 404.1520a (2008). See generally Thomas P. Harding, *Psychiatric Disability and Clinical Decision Making: The Impact of Judgment Error and Bias*, 24 Clinical Psych. Rev. 707 (2004), <https://doi.org/10.1016/j.cpr.2004.06.003>; Cille Kennedy, *SSA’s Disability Determination of Mental Impairments: A Review Toward an Agenda for Research*, in *The Dynamics of Disability: Measuring and Monitoring Disability for Social Security Programs* 241 (Gooloo S. Wunderlich et al. eds., 2002); Dan B. Dobbs, *The Law of Torts* 1048–53, 1087–1110 (2000); Andrew W. Kane & Joel A. Dvoskin, *Evaluation for Personal Injury Claims* (2011); Lisa Drago Piechowski, *Evaluation of Workplace Disability* (2011).

168. See discussion in section titled “Functional Impairment Due to Mental Disorders” above.

due to mental disorders that are likely to occur over time. In contrast to the structured assessment tools that assist in the prediction and management of future violence, no instruments have been developed and validated to predict future functional status.¹⁶⁹ Such predictions are complicated by the need for simultaneous estimates of several parameters that affect long-term functional outcomes: variables intrinsic to the person (e.g., symptomatic fluctuation, changes in motivation to work), variables that relate to the environment (e.g., divorce, availability of new categories of jobs), and responses to treatment.¹⁷⁰ Research exists that is aimed at identifying variables associated with some types of future functional impairment, but it is largely focused on progressive disorders (e.g., Alzheimer's disease), and even in those circumstances the accuracy of the predictions of forensic evaluators has not been determined.¹⁷¹ Acknowledgment of the uncertainties inherent in these predictions is therefore essential for experts undertaking this task.

Risk-Need-Responsivity (RNR) Assessments

Two related questions have also been addressed in the risk-assessment literature: whether there are risk-relevant needs (risk factors that are potentially changeable and, if changed, would affect the overall appraisal of risk), and how an individual might respond to interventions targeting these needs (responsivity). Risk-need-responsivity (RNR) was originally proposed as an approach to assessing the risk and rehabilitation needs of individuals who had already been convicted and sentenced to incarceration.¹⁷² Since its inception, however, it has influenced other approaches to risk assessment in legal contexts, in part because it can address questions that are not necessarily answered by predictions but may be of interest to the judge. Appraisal that focuses on RNR can be distinguished from assessment focusing only on prediction: Although both strive to identify the likelihood that the outcome of interest will occur, RNR considers the additional questions of whether that likelihood can be changed through planned intervention—and whether such intervention is likely to have the desired

169. Such instruments can be very helpful to forensic clinicians, as they summarize the available scientific evidence, provide procedures that apply that evidence, guide the conclusions of the evaluator by structuring the gathering of evidence and translating that raw information (using scoring and/or professional judgment), and provide a manual summarizing the relevant information in a single source.

170. See discussion in section titled “Prediction of Responses to Treatment” below.

171. See, e.g., Roy Martin et al., *Declining Financial Capacity in Patients with Mild Alzheimer Disease: A One-Year Longitudinal Study*, 16 Am. J. Geriatric Psychiatry 209 (2008), <https://doi.org/10.1097/JGP.0b013e318157cb00>.

172. See Donald A. Andrews et al., *Classification for Effective Rehabilitation: Rediscovering Psychology*, 17 Crim. Just. Behav. 19 (1990), <https://doi.org/10.1177/0093854890017001004>.

risk-reducing effect. Depending on the questions associated with the legal decision, one of these two approaches might be a better fit. An RNR appraisal also focuses on criminal offending generally rather than violence specifically.

Appraising Risk

Risk factors can be variables that do not change through planned intervention (called static risk factors, e.g., sex or age) or variables that have the potential to change through such intervention (called dynamic risk factors, e.g., substance use disorder). Predictive approaches can use either static or dynamic risk factors, although the former change little in their level of appraised risk over time. RNR, by contrast, uses mostly dynamic risk factors—also termed needs. Appraisals using both static and dynamic risk factors yield an overall risk category that may change over time and with the delivery of planned rehabilitative interventions. It was once accepted that static risk factors provided a stronger basis for accurate prediction, but current evidence suggests that this superiority is smaller than was once believed and may not exist.¹⁷³ In other words, the risk of future violence or other criminal offending is subject to change for many individuals, and the risk assessment that is provided should reflect that.

Appraising Need

Areas of need differ across different justice-involved populations. For individuals who have been termed general offenders, who are not selected for particular characteristics (e.g., clinical characteristics such as symptoms of severe mental illness; offending characteristics such as being charged with a specific type of offense), risk-relevant needs have been identified in the following areas: criminal thinking, antisocial personality, antisocial peers, substance misuse, family, education/employment, and leisure time.¹⁷⁴ When combined with the static historical variable of antisocial conduct, these risk factors have been called the central eight.¹⁷⁵ Among individuals with severe mental illness, research has identified two of these needs (severe substance use/dependence and criminal thinking) as particularly important, as well as some other dimensions that are not among

173. See Mara J. Eisenbert et al., *Static and Dynamic Predictors of General and Violent Criminal Offense Recidivism in the Forensic Outpatient Population: A Meta-Analysis*, 46 *Crim. Just. Behav.* 732 (2019), <https://doi.org/10.1177/0093854819826109>.

174. See James Bonta & Donald A. Andrews, *The Psychology of Criminal Conduct* (6th ed. 2017).

175. See, e.g., David DeMatteo et al., *Problem-Solving Courts and the Criminal Justice System* 34 (2019).

the original central eight (self-management, social skills, and life skills). The identification of risk-relevant needs varies somewhat according to population (e.g., general offenders, severely mentally ill individuals, specialized offenders such as those charged with sex offenses) and outcome of interest (e.g., violence, serious violence, sex offending, any offending).

Appraising Responsivity

When considering the need for risk-reducing interventions in different areas, a question arises: How are individuals likely to respond to such interventions? This is the key question in appraising responsivity, the third domain of risk-need-responsivity. Two types of responsivity have been identified: general and specific.¹⁷⁶ General responsivity refers to the importance of using interventions that have been demonstrated through research to be effective. This means that interventions with demonstrated effectiveness in reducing antisocial behavior (e.g., medication, therapeutic communities, cognitive behavioral therapy) should be preferred, while those without such evidence (e.g., psychoanalysis) should be viewed with skepticism if proposed for this purpose. Specific responsivity underscores the importance of matching the demands of the intervention with the capacities and learning style of the individual. For example, an individual with identified intellectual limitations in verbal skills and verbal memory should not receive an intervention that depends on the individual's ability to describe and recall complex ideas; rather, the focus should be on developing and strengthening adaptive behavior.

RNR may be particularly relevant at sentencing. Toward that end, a set of national guidelines for postconviction risk and need assessment has been developed with the assistance of numerous experts in the field.¹⁷⁷

Treatment of Mental Disorders

The nature of available treatments for mental disorders, the probability that the treatments will be effective, the side effects they may induce, and the existence of alternatives are likely to be material to a variety of legal cases. In criminal proceedings, for example, the continued confinement of a defendant in a psychiatric hospital on the basis of incompetence to stand trial will be based in part on the probability that treatment will restore capacity;¹⁷⁸ involuntary treatment of

176. See Bonta & Andrews, *supra* note 174.

177. Sarah L. Desmarais et al., *Advancing Fairness and Transparency: National Guidelines for Post-Conviction Risk and Needs Assessment* (2022).

178. See *Jackson v. Indiana*, 406 U.S. 715 (1972).

the defendant will turn on a number of factors, including the likelihood of treatment success, treatment side effects, and the potential for side effects to impair the defendant's defense.¹⁷⁹ Decisions about probation and parole of mentally disordered offenders may also depend on the likelihood that symptoms will remain in check, and courts may order ongoing treatment as a condition of release.¹⁸⁰ Among the civil cases for which treatment-related questions will be at issue are liability claims for malpractice and failure to protect third parties from patient violence, claims involving emotional harms (e.g., in calculating the cost of future care), and issues related to the deprivation of rights of prisoners in correctional facilities to have adequate mental health treatment.¹⁸¹ Treatment of mental disorders today offers multiple options for most disorders, often with different levels of likely effectiveness and varying side-effect profiles. Planning treatment has become an increasingly complex task.

Treatment with Medication

The past seventy years have seen the ongoing introduction of new medications for the treatment of mental disorders. Currently, medications are a mainstay in the treatment of schizophrenia and bipolar disorder; it is a rare patient who can be treated successfully for these disorders without medication as part of the treatment plan.¹⁸² Medications are also used commonly to treat and prevent the recurrence of depression, anxiety disorders, attention-deficit/hyperactivity disorder (ADHD), and a large number of other conditions.¹⁸³ The field of psychopharmacology, as the treatment of mental disorders with medications is known, has become a complex and challenging part of psychiatric practice.¹⁸⁴

Psychological Treatments

Although medications are a mainstay for treatment of serious mental disorders, a variety of psychological treatments may be important as either primary or adjunctive treatments.

179. *See* *Sell v. United States*, 539 U.S. 166 (2003).

180. *See, e.g., United States v. Holman*, 532 F.3d 284 (4th Cir. 2008).

181. For an overview of the considerable body of case law on this issue, see Michael L. Perlin, 2 *Mental Disability Law* § 11–4.3 (2005).

182. This applies to a range of criminal and civil questions involving rehabilitation of such individuals (e.g., competence to stand trial, hospitalization, sentencing, civil injury).

183. *See generally* Alan F. Schatzberg & Charles DeBattista, *Schatzberg's Manual of Clinical Psychopharmacology* (9th ed. 2019); Benjamin J. Sadock, Norma Sussman & Virginia A. Sadock, Kaplan & Sadock's *Pocket Handbook of Psychiatric Drug Treatment* (7th ed. 2019).

184. *See generally* Herbert Mwebe, *Psychopharmacology: A Mental Health Professional's Guide to Commonly Used Medications* (2018).

Behavioral Therapy and Cognitive Behavioral Therapy

The goal of behavioral therapy is to reduce maladaptive behaviors (e.g., substance use, impulsivity) by reinforcing desirable behaviors and eliminating unwanted behaviors.¹⁸⁵ Behavioral therapy is based on the idea that people learn behaviors based on their environment and experiences; it seeks to help people understand, among other things, the antecedents and consequences of their behaviors, thereby motivating them to change unwanted behaviors.¹⁸⁶ Cognitive behavioral therapy (CBT) is based on the idea that patterns of thought determine how one feels and behaves.¹⁸⁷ CBT is generally shorter-term (weeks to months), highly structured, and focused on helping patients recognize and control maladaptive patterns of thinking. Patients are often given homework assignments to complete between sessions. A strong database supports its use in anxiety disorders, depression (where it can be as effective as medications and may be more likely to prevent relapse), and for control of some psychotic symptoms, and its use is steadily being extended to additional conditions.¹⁸⁸

Specialized Treatments

Given the effectiveness of CBT, several specialized forms of CBT have been developed for certain populations (e.g., individuals who commit sex offenses) based on a model that is often termed relapse prevention, which teaches patients to recognize and avoid situations that are likely to lead to recidivism.¹⁸⁹ Dialectical behavior therapy (DBT) is an offshoot of CBT that has shown success with patients with borderline personality disorders, a disorder otherwise difficult to treat.¹⁹⁰ Mindfulness-based interventions, which can range from meditation to more formal interventions that incorporate CBT techniques, are designed to increase intentional attention and cultivate different strategies for dealing with distressing thoughts and emotions,¹⁹¹ with research suggesting that they are effective for certain psychiatric disorders.¹⁹²

185. See Joseph Wolpe, *The Practice of Behavior Therapy* (1969).

186. See *id.*

187. Aaron T. Beck & Cory F. Newman, *Cognitive Therapy*, in Sadock et al., *supra* note 104, at 2595; Judith S. Beck, *Cognitive Therapy: Basics and Beyond* (1995).

188. Andrew C. Butler et al., *The Empirical Status of Cognitive-Behavioral Therapy: A Review of Meta-Analyses*, 26 *Clinical Psych. Rev.* 17 (2006), <https://doi.org/10.1016/j.cpr.2005.07.003>.

189. See, e.g., D. Richard Laws, *Relapse Prevention with Sex Offenders* (1989).

190. M. Zachary Rosenthal & Thomas R. Lynch, *Dialectical Behavior Therapy*, in Sadock et al., *supra* note 104, at 2619.

191. See Benjamin G. Shapero et al., *Mindfulness-Based Interventions in Psychiatry*, 16 *Focus* 32 (2018), <https://doi.org/10.1176/appi.focus.20170039>.

192. See Simon B. Goldberg et al., *The Empirical Status of Mindfulness-Based Interventions: A Systematic Review of 44 Meta-Analyses of Randomized Controlled Trials*, 17 *Persp. Psych. Sci.* 108 (2022), <https://doi.org/10.1177/1745691620968771>.

Treatment of Functional Impairments

Control of positive symptoms does not necessarily address deficits in function, particularly in the psychotic disorders. What may be required are techniques that focus on functional difficulties per se. For example, given that schizophrenia often affects the ability to function socially and occupationally, persons with the disorder may need to be taught how to interact with other people, an approach known as social-skills therapy.¹⁹³ Occupational therapy can provide them with a graded introduction (or reintroduction) to the workplace, with patients taught how to maintain focus and deal with the demands of the work setting.¹⁹⁴ More focal impairments can be addressed, as well. Thus, defendants found incompetent to stand trial can be taught about the nature of the courtroom, the role of key legal players, and the expectations they must meet to be found competent to proceed. Studies of such programs have shown higher rates of restoration of competence than occurs with treatment of the primary disorder alone.¹⁹⁵ Comparable programs are available for anger management,¹⁹⁶ control of spousal abuse,¹⁹⁷ and training in parenting skills,¹⁹⁸ among other areas of function that are often the target of legal proceedings.

Prediction of Responses to Treatment

In a number of legal contexts, experts are called on to anticipate the response to treatment of persons with mental disorders. These projections are difficult to make because of several parameters inherently challenging to predict:

193. Alex Kopelowicz, Robert Paul Liberman & Roberto Zarate, *Recent Advances in Social Skills Training for Schizophrenia*, 32 *Schizophrenia Bull.* S12 (2006), <https://doi.org/10.1093/schbul/sbl023>.

194. See generally Jennifer Creek, *Occupational Therapy and Mental Health: Principles, Skills and Practice* (2002).

195. See, e.g., Alex M. Siegel & Miriam Elwork, *Treating Incompetence to Stand Trial*, 14 *L. & Hum. Behav.* 57 (1990), <https://doi.org/10.1007/BF01055789>; Kirk Heilbrun et al., *Jackson-Based Restorability to Competence to Stand Trial: Critical Analysis and Recommendations*, 27 *Psych. Pub. Pol. & L.* 370 (2021), <https://doi.org/10.1037/law0000307>; Barry W. Wall et al., *Restoration of Competency to Stand Trial: A Training Program for Persons with Mental Retardation*, 31 *J. Am. Acad. Psychiatry & L.* 189 (2003).

196. Raymond DiGiuseppe & Raymond C. Tafrate, *Anger Treatment for Adults: A Meta-Analytic Review*, 10 *Clinical Psych.: Sci. & Prac.* 70 (2006), <https://doi.org/10.1093/clipsy.10.1.79>.

197. Julia C. Babcock et al., *Does Batterers' Treatment Work? A Meta-Analytic Review of Domestic Violence Treatment*, 23 *Clinical Psych. Rev.* 1023 (2004), <https://doi.org/10.1016/j.cpr.2002.07.001>. Note that in contrast to anger management and parenting training, the data on the efficacy of treatment for batterers indicate that effects are limited at best.

198. Kathryn M. Bigelow & John R. Lutzker, *Training Parents Reported for or at Risk for Child Abuse and Neglect to Identify and Treat Their Children's Illnesses*, 15 *J. Fam. Violence* 311 (2000), <https://doi.org/10.1023/A:1007550028684>.

- *Effectiveness of Treatment.* Even highly effective treatments do not work in all cases, and when they do work, they may provide varying levels of relief.¹⁹⁹
- *Adherence.* Treatment has no chance of being effective if a person declines to pursue or to continue it, a particular issue in cases where the court lacks control over the person's future behavior.²⁰⁰ Tolerability of side effects may play an important role in these decisions.
- *Fluctuations in the Course and Responsiveness of the Disorder.* Many mental disorders are chronic and tend to wax and wane in intensity. Although adjustments in treatment can sometimes bring more severe symptoms under good control, that is not always possible. For reasons that are not understood, previously responsive disorders may also become resistant to the therapeutic effects of medication.²⁰¹
- *Environmental Conditions.* Unpredictable stresses in a person's life may exacerbate symptoms, reduce the effectiveness of treatment, or lead to diminished adherence.

However, given that estimates sometimes must be made of probable treatment effects, there are several indicators to which clinicians can turn.²⁰² Previous treatment response is the best predictor of future response; it is likely, for example, that someone whose previous delusions have rapidly resolved with antipsychotic medication will have a similar response in the future. In the absence of a documented history of successful treatment, estimates should be

199. For example, only 45% to 60% of patients receiving antidepressant medication for uncomplicated major depression show clinically significant responses to the first medication they receive, and of those who fail to respond, a similar percentage will respond positively to a second medication. Rates of response in unselected populations of patients with depression are lower. Madhukar H. Trivedi et al., *Evaluation of Outcomes with Citalopram for Depression Using Measurement-Based Care in STAR*D: Implications for Clinical Practice*, 163 Am. J. Psychiatry 28 (2006), <https://doi.org/10.1176/appi.ajp.163.1.28>.

200. Rates of nonadherence to medications among patients with psychiatric disorders are in the range of 50% or more. Although these figures are perhaps somewhat higher than those seen in other chronic conditions, long-term treatment with medication in general is marked by high rates of noncompliance with prescribed medications. Lars Osterberg & Terrence Blaschke, *Adherence to Medication*, 353 New Eng. J. Med. 487 (2005), <https://doi.org/10.1056/NEJMra050100>.

201. So-called poop-out during treatment of depression is a commonly encountered example. See, e.g., Sarah E. Byrne & Anthony J. Rothschild, *Loss of Antidepressant Efficacy During Maintenance Therapy: Possible Mechanisms and Treatments*, 59 J. Clinical Psychiatry 279 (1998), <https://doi.org/10.4088/jcp.v59n0602>.

202. For predictors of response to treatment for depression, see, for example, Stuart M. Sotsky et al., *Patient Predictors of Response to Psychotherapy and Pharmacotherapy: Findings in the NIMH Treatment of Depression Collaborative Research Program*, 148 Am. J. Psychiatry 997 (1991), <https://doi.org/10.1176/foc.4.2.278>; for predictors of response to treatment for schizophrenia, see Delbert G. Robinson et al., *Predictors of Treatment Response from a First Episode of Schizophrenia or Schizoaffective Disorder*, 156 Am. J. Psychiatry 544 (1999), <https://doi.org/10.1176/ajp.156.4.544>.

based on evidence indicating base rates of response for the person's disorder, along with any specific prognostic factors present in the person's case (e.g., a schizophrenic disorder that develops slowly over many years and that is associated with gradual functional decline generally has a poorer prognosis than one with rapid onset and good premorbid functioning). To some extent, however, there is always uncertainty associated with these predictions.

Treatment to Reduce Risk of Reoffending

A number of treatments have been developed specifically to reduce the risk of reoffending among justice-involved individuals. As a starting point, it is important to note that mental illness has a modest and indirect impact on criminal offending, and although mental illness can lead to increased contact with law enforcement (due, for example, to odd or disruptive behavior), mental illness is not a strong risk factor for violent behavior.²⁰³ Of note, mental illness is not one of the key risk factors for offending in the RNR model discussed earlier in this reference guide. Although mental illness is weakly associated with criminal offending, it can affect other risk factors; for example, offenders who are mentally ill may be disproportionately exposed to other risk factors that can increase reoffense risk, such as unstable housing, unemployment, substance misuse, criminal peers, and victimization.²⁰⁴ Several interventions have been designed to reduce the risk of reoffending.

Treatment with Medications

A variety of medications have some impact on reducing impulsive and/or aggressive behavior.²⁰⁵ The use of antipsychotic medications among people with schizophrenia who are not justice-involved has been shown to have a minimal

203. See D. A. Andrews, James Bonta & J. Stephen Wormith, *The Recent Past and Near Future of Risk and/or Need Assessment*, 52 *Crime & Delinquency* 7 (2006), <https://doi.org/10.1177/0011128705281756>; Jacques Baillargeon et al., *Psychiatric Disorders and Repeat Incarcerations: The Revolving Prison Door*, 166 *Am. J. Psychiatry* 103 (2009), <https://doi.org/10.1176/appi.ajp.2008.08030416>; Kristin G. Cloyes et al., *Time to Prison Return for Offenders with Serious Mental Illness: A Survival Analysis*, 37 *Crim. Just. Behav.* 175 (2010), <https://doi.org/10.1177/0093854809354370>.

204. See Jennifer L. Skeem, Sarah Manchak & Jillian K. Peterson, *Correctional Policy for Offenders with Mental Illness: Creating a New Paradigm for Recidivism Reduction*, 35 *L. & Hum. Behav.* 110 (2011), <https://doi.org/10.1007/s10979-010-9223-7>.

205. See Peter W. Schofield et al., *Pharmacotherapy to Reduce Violent Offending? Offenders Might Be Interested*, 53 *Austl. & N.Z. J. Psychiatry* 697 (2019), <https://doi.org/10.1177/0004867419835937>.

effect in reducing violence, largely because schizophrenia is not a major risk factor for violent behavior,²⁰⁶ although it may be useful when a person's psychotic symptoms are linked directly to their violence. Some other medications, including selective serotonin reuptake inhibitors (SSRIs), have shown some promise in reducing (albeit modestly) aggressive and antisocial behavior.²⁰⁷ The use of any medication to reduce the risk of reoffending among those without mental illness raises ethical concerns that must be balanced against the desired objective.

Psychological Treatments

There are several interventions that target empirically supported risk factors for offending and violence. For example, interventions that address the risk factors identified in the RNR model can be effective in reducing the incidence of antisocial behavior. Specifically, interventions that teach problem-solving skills, anger management, prosocial thinking, and coping skills can reduce antisocial behavior, as can educational/vocational training and interventions designed to reduce antisocial associations, improve conflict-resolution skills, reduce substance use, and increase involvement in prosocial pursuits.²⁰⁸

In recent years, manualized treatments designed to reduce recidivism have been developed. Changing Lives and Changing Outcomes (CLCO) is a manualized treatment designed for justice-involved individuals with serious mental illness, including those with a dual diagnosis of mental illness and substance use.²⁰⁹ CLCO is a CBT-based approach that helps justice-involved individuals to understand and cope with their disorders, identify and challenge antisocial thought patterns, learn basic social skills, abstain from substance use, and proactively manage their disorders.²¹⁰ The dual focus on risk factors (or criminogenic needs) and relevant clinical needs provides a strong foundation for reducing risk. CLCO is associated with higher treatment engagement, clinically meaningful reductions in mental-disorder symptoms and criminal thinking, increased

206. See Jeffrey W. Swanson et al., *Comparison of Antipsychotic Medication Effects on Reducing Violence in People with Schizophrenia*, 193 Brit. J. Psychiatry 37 (2008), <https://doi.org/10.1192/bjp.bp.107.042630>.

207. See Tony Butler et al., *Reducing Impulsivity in Repeat Violent Offenders: An Open Label Trial of a Selective Serotonin Reuptake Inhibitor*, 44 Austl. & N.Z. J. Psychiatry 1137 (2010), <https://doi.org/10.3109/00048674.2010.525216>.

208. See Bonta & Andrews, *supra* note 174.

209. See Robert D. Morgan, Daryl G. Kroner & Jeremy F. Mills, *A Treatment Manual for Justice-Involved Persons with Mental Illness: Changing Lives and Changing Outcomes* (2017).

210. See *id.*

knowledge of treatment-related information, and increased medication adherence.²¹¹

Limitations of Mental Health Evidence

There are certain limitations to mental health evidence that there may not be with other types of scientific evidence. Both retrospective assessments of past mental states and prospective estimates of future behavior depend on estimates of variables that are inherently difficult to make with a high degree of certainty. Even contemporaneous assessments of functional abilities depend in part on the evaluatee's self-report of such difficult-to-measure attributes as distress, motivation, and judgment. Where empirically validated assessment tools are used, there are the usual concerns about measurement error. Two other key issues relate to ultimate-issue testimony and the standard to which expert testimony is typically held—that is, reasonable degree of certainty.

Ultimate-Issue Testimony

Whether mental health experts should testify—or be permitted to testify—to the ultimate legal issue in a case has been a subject of long-standing controversy.²¹² The question arises, for example, in criminal cases where experts often have commented directly on whether a defendant is competent to stand trial or whether the legal standard for insanity has been met.²¹³ Similar issues can arise in civil settings, in which experts may be asked to testify directly about a person's capacity to manage affairs or to serve as a custodial parent, or regarding whether a person was competent to sign a contract at an earlier point in time.²¹⁴

Attorneys and judges are often proponents of ultimate-issue testimony. They may be concerned that an expert who provides a clinical formulation without tying it directly to the ultimate legal issue will confuse jurors, who may

211. See, e.g., Monika Gaspar et al., *Therapeutic Outcomes of Changing Lives and Changing Outcomes for Male and Female Justice Involved Persons with Mental Illness*, 46 *Crim. J. & Behav.* 1678 (2019), <https://doi.org/10.1177/0093854819879743>; Stephanie A. Van Horn et al., *Changing Lives and Changing Outcomes: “What Works” in an Intervention for Justice-Involved Persons with Mental Illness*, 16 *Psych. Servs.* 693 (2019), <https://doi.org/10.1037/ser0000248>.

212. See Fed. R. Evid. 704; Anne Lawson Braswell, *Resurrection of the Ultimate Issue Rule: Federal Rule of Evidence 704(b) and the Insanity Defense*, 72 *Cornell L. Rev.* 620 (1987); Christopher Slobogin, *The “Ultimate Issue” Issue*, 7 *Behav. Sci. & L.* 259 (1989).

213. But see discussion below regarding the current prohibition on this practice in federal courts.

214. See Restatement (Second) of Contracts § 15.

be unable to discern the connection on their own.²¹⁵ Many experts share similar concerns or worry that mental health issues will simply be ignored if their relevance to the legal question at hand is not made clear; they further note that courts have applied such rules erratically.²¹⁶ They counter concerns about such testimony having an undue impact on jurors' deliberations by noting that jurors appear to be little influenced by whether ultimate-issue testimony is offered by an expert.²¹⁷

In addition to the argument that ultimate-issue testimony invades the province of the decision maker,²¹⁸ opponents often point to the legal and moral nature of the question of whether someone is, for example, criminally responsible.²¹⁹ Although mental health expertise may be helpful in determining the person's mental state at the relevant time, determining whether the resulting impairment was sufficient to negate responsibility requires the application of the relevant legal standard and a moral judgment of the fairness or unfairness of punishing a person for the behavior. Psychiatrists and psychologists have no particular expertise on legal or moral issues; hence, opponents of ultimate-issue testimony urge that mental health experts should not be permitted to speak to those issues. Such preclusion may also reduce the much bemoaned "battle of the experts," because a good deal of disagreement may derive from views of how data from the evaluation should be applied to the ultimate legal question rather than from differences regarding the person's mental state. Although testimony on the ultimate legal issue is now barred in federal courts in insanity-defense cases (18 U.S.C. § 17; Federal Rule of Evidence 704(b)), ultimate-issue testimony is permitted in many states, and, even in federal jurisdictions, it may be offered in other cases.²²⁰

215. Ralph Slovenko, *Commentary: Deceptions to the Rule on Ultimate Issue Testimony*, 34 J. Am. Acad. Psychiatry & L. 22 (2006). Cases involving issues related to medical causation can be particularly confusing for fact finders if experts are not permitted to address the ultimate legal issue. See Slobogin et al., *supra* note 1, at 624.

216. Alec Buchanan, *Psychiatric Evidence on the Ultimate Issue*, 34 J. Am. Acad. Psychiatry & L. 14 (2006).

217. Solomon M. Fulero & Norman J. Finkel, *Barring Ultimate Issue Testimony: An "Insane" Rule?* 15 L. & Hum. Behav. 495 (1991), <https://doi.org/10.1007/BF01650291>.

218. Insanity Defense Workgroup, *American Psychiatric Association Position on the Insanity Defense*, 140 Am. J. Psychiatry 681, 686 (1983), <https://doi.org/10.1176/ajp.140.6.681>; American Bar Association, ABA Criminal Justice Standards: Mental Health, Standard 7–6.6 (1984). See also Grisso, *supra* note 3, at 208; Fulero & Finkel, *supra* note 217, at 496.

219. Mark S. Brodin, *Behavioral Science Evidence in the Age of Daubert: Reflections of a Skeptic*, 73 U. Cin. L. Rev. 867 (2005); Michele Cotton, *A Foolish Consistency: Keeping Determinism Out of the Criminal Law*, 15 B.U. Pub. Int. L. J. 1, 21–23 (2005); Ric Simmons, *Conquering the Province of the Jury: Expert Testimony & the Professionalization of Fact-Finding*, 74 U. Cin. L. Rev. 1013 (2006).

220. Fed. R. Evid. 704. Pennsylvania's law offers a typical formulation: "Testimony in the form of an opinion or inference otherwise admissible is not objectionable because it embraces an ultimate issue to be decided by the trier of fact." Pa. R. Evid. 704.

“Reasonable Degree of Certainty”

How much confidence do experts need in their opinion before they can state it in court? In many jurisdictions, experts are required to state their opinions to a “reasonable degree of certainty” within the standards of the relevant field, although such terminology is not required by the Federal Rules of Evidence, the Federal Rules of Criminal Procedure, or the Federal Rules of Civil Procedure. Further, such language is not mandated by *Frye*,²²¹ *Daubert*,²²² or *Kumho Tire*,²²³ and it is not clearly defined by any case law.²²⁴ Some state courts have held that the phrase is required as a prerequisite to the admissibility of expert evidence regardless of how certain an expert is in the opinion,²²⁵ while other state courts require a high degree of certainty and use of the phrase.²²⁶ The high courts of at least two states have indicated that the phrase is not required.²²⁷

There are several concerns with requiring use of the phrase “reasonable degree of [scientific, medical, psychological, etc.] certainty.” The phrase lacks a common definition across scientific disciplines and jurisdictions, and it has alternatively been defined or equated with more likely than not, more probable than not, 51% likelihood, preponderance of the evidence, clear and convincing evidence, and beyond a reasonable doubt.²²⁸ Such phrasing may also lend more credence to the testimony than is actually deserved, particularly when the testimony is presented in probabilistic terms, and it can be misleading to the factfinder about the level of objectivity involved in the analysis, its scientific reliability and limitations, and the ability of the expert to reach an

221. *Frye v. United States*, 293 F. 1013 (1923).

222. *Daubert v. Merrell Dow Pharms., Inc.*, 509 U.S. 579 (1993).

223. *Kumho Tire Co. v. Carmichael*, 526 U.S. 137 (1999).

224. The phrase “reasonable degree of certainty” was first applied to scientific evidence in 1935 when a witness was asked if he could determine, with reasonable scientific certainty, the cause of the capsizing of a boat. See *Herbst v. Levy*, 279 Ill. App. 353 (Ill. App. Ct. 1935). In that case, such phrasing was not a legal mandate but the stylistic approach of the lawyer. The phrase was linked to admissibility in 1969 in *Twin City Plaza, Inc. v. Central Surety & Insurance Corp.*, 409 F.2d 1195 (8th Cir. 1969), in which the U.S. Court of Appeals for the Eighth Circuit held that a witness’s testimony should only be admitted if, inter alia, it was based on reasonable scientific certainty.

225. In *Bertram v. Wunning*, 385 S.W.2d 803 (Mo. Ct. App. 1967), a medical expert’s testimony that he was 90% certain in his opinion was held to be insufficient because he refused to state that his opinion was offered to a reasonable degree of medical certainty.

226. In *Griffin v. Univ. of Pittsburgh Med. Center-Braddock Hosp.*, 950 A.2d 996 (Pa. Sup. Ct. 2008), an expert’s opinion, despite being offered to a reasonable degree of medical certainty, was held to be insufficient because the expert acknowledged being only 51% sure of his opinion.

227. See *Hawaii v. DeLeon*, 319 P.3d 282 (Haw. 2014); *Nebraska v. Johnson*, 290 Neb. 862 (2015).

228. See Gutheil & Appelbaum, *supra* note 69.

individualized conclusion. In short, such phrasing, without proper definition and guidance, is a relic of custom and practice that can invite confusion.²²⁹

Evaluating Evidence from Mental Health Experts

Up to this point, we have considered the kind of evidence that is likely to be offered by mental health experts and some of the challenges that such testimony presents. The remainder of the reference guide addresses the factors that should enter into the consideration of the value and impact of such testimony.

What Are the Qualifications of the Expert?

The appropriate qualifications of a mental health professional whose testimony is proffered will depend on the nature of the evidence to be presented. However, there are a number of relevant parameters that can be identified.

Training

Most mental health expert testimony is given by psychiatrists or doctoral-level clinical psychologists. Given the differences in the education and training of each profession, their testimony is not necessarily interchangeable. As a rule, psychiatrists are prepared by their training to speak to the diagnosis of mental disorders, including medical issues that may play a role in a particular case, and to treatment approaches, including psychopharmacological treatment.²³⁰ They

229. In 2016, the National Commission on Forensic Science made several recommendations to the U.S. attorney general regarding use of the phrase: (1) the attorney general should direct all attorneys appearing on behalf of the Department of Justice (DOJ) (a) to forego use of such phrases when presenting forensic testimony unless required by judicial authority as a condition of admissibility for a witness's opinion, and (b) to assert a legal position that such terminology is not required and is misleading; (2) the attorney general should direct all forensic experts employed by the DOJ not to use such language in reports or couch their testimony in such terms unless directed to do so by judicial authority; and (3) the attorney general should develop appropriate language that may be used by experts when reporting or testifying about results or findings based on observations of evidence and data derived from evidence. The full text can be found at <https://perma.cc/7GPX-C4RX>.

230. See discussion of psychiatrists' training in section titled "Psychiatrists" above.

should be capable of testifying, within the limits of existing knowledge and the information available to them, regarding the impact of a disorder on a person's behavior and functional abilities. Psychologists' training, in contrast, may provide deeper knowledge of the theoretical and experimental bases for understanding the function of the mind and the influences on behavior, both normal and abnormal.²³¹ As a general matter, doctoral-level clinical psychologists will be prepared by their training to provide evidence regarding diagnosis and psychotherapeutic treatment of mental disorders, the results of psychological and neuropsychological testing, and the roots of normal and abnormal behavior. More specifically, both psychiatrists and psychologists can subspecialize in forensic psychiatry and forensic psychology, respectively. These specializations include training in mental health evaluations conducted in the course of litigation, for the purpose of informing the court and/or helping the retaining attorney to present their case. Psychiatrists may formalize this specialization through a forensic psychiatry fellowship.²³² Psychologists can receive specialty training as part of their predoctoral or internship work and can also do a postdoctoral forensic fellowship.²³³

Although the core elements of training in psychiatry and psychology may be similar across training programs, the variability is substantial.²³⁴ Moreover, variation in subspecialty (in psychiatry) or specialty (in psychology) training—for example, in geriatric psychiatry or neuropsychology—contributes to further differentiation among experts. Thus, it may be necessary to inquire into the specific training of an expert. This is particularly true when an expert is testifying about topics that would ordinarily fall outside disciplinary boundaries—for example, a psychiatrist discussing the results of psychological testing or a psychologist offering evidence regarding the effect of medication on a person's behavior. The same is true for experts who are testifying beyond the range of their specialty or subspecialty training. In addition, in recent years expert testimony on mental health issues at times has been admitted from nonpsychiatric physicians and mental health professionals of other disciplines.²³⁵ These include

231. See discussion of psychologists' training in section titled "Psychologists" above.

232. American Academy of Psychiatry and the Law, Directory of Forensic Psychiatry Fellowships, <https://perma.cc/M9BG-ZEK3>.

233. See, e.g., David DeMatteo et al., *Becoming a Forensic Psychologist* (2019).

234. See, e.g., Helena Hansen, Joel Braslow & Robert M. Rohrbaugh, *From Cultural to Structural Competency—Training Psychiatry Residents to Act on Social Determinants of Health and Institutional Racism*, 75 JAMA Psychiatry 117 (2018), <https://doi.org/10.1001/jamapsychiatry.2017.3894>; David A. Ross, Michael J. Travis & Melissa R. Arbuckle, *The Future of Psychiatry as Clinical Neuroscience: Why Not Now?*, 72 JAMA Psychiatry 413 (2015), <https://doi.org/10.1011/jamapsychiatry.2014.3199>; *The Oxford Handbook of Education and Training in Professional Psychology* (Nadine J. Kaslow & W. Brad Johnson eds., 2014).

235. *Campbell v. Metro. Prop. & Cas. Ins. Co.*, 239 F.3d 179 (2d Cir. 2001) (professor of pediatrics with substantial relevant publications found qualified to testify on neurological injuries resulting from lead paint exposure); *Carroll v. Otis Elevator Co.*, 896 F.2d 210 (7th Cir. 1990)

social work and nursing and could include other disciplines as well (e.g., master's-level psychologists, marriage and family therapists, physician assistants). One reasonable suggestion for qualifying experts involves applying a standard like that set forth in Rule 702²³⁶ (probable assistance to the trier of fact), in which the judge could consider both general training and more specific influences, such as specialized forensic training and/or experience, in deciding whether to qualify a given professional as an expert.²³⁷ These additional influences are considered in the following sections.

Experience

Experience is relevant to the qualifications of mental health experts in at least two ways. First, as the Federal Rules of Evidence recognize, experience may substitute for training as a basis for concluding that a witness has special expertise.²³⁸ Some experts in forensic psychiatry and forensic psychology, for example, lack formal training in conducting evaluations of the sort provided in forensic fellowships, because such training programs were not widely available at the time when senior forensic psychiatrists and psychologists were trained. In addition, formal training is difficult to obtain in some substantive areas of clinical psychiatry and psychology. Some professionals who acquire special knowledge about particular mental disorders will do so by pursuing their interest through reading, following the literature, continuing professional education, and through clinical contact with patients with the disorders, as opposed to formal training at the residency, fellowship, and/or predoctoral levels. Thus, experience must sometimes be relied upon as a stand-in for more formal training in some areas.

(experimental psychologist found qualified to give expert testimony on likelihood that product design would cause children to press escalator's emergency stop button); *United States v. Withorn*, 204 F.3d 790 (8th Cir. 2000) (trial court properly admitted testimony from midwife on alleged sexual assault on basis of bachelor's degree, some postgraduate work, and clinical experience). *But see* *United States v. Moses*, 137 F.3d 894 (8th Cir. 1998) (social worker lacked expertise to opine that victim of alleged child abuse would suffer trauma from facing the accused abuser in the courtroom).

236. "If scientific, technical, or other specialized knowledge will assist the trier of fact to understand the evidence or to determine a fact in issue, a witness qualified as an expert by knowledge, skill, experience, training, or education, may testify thereto in the form of an opinion or otherwise, if (1) the testimony is based upon sufficient facts or data, (2) the testimony is the product of reliable principles and methods, and (3) the witness has applied the principles and methods reliably to the facts of the case." Fed. R. Evid. 702 (2000).

237. See Melton et al., *supra* note 1.

238. This may apply particularly to psychiatrists and psychologists who have not received formal forensic training but who have experience with individuals similar to the criminal defendant or civil litigant before the court.

The second way in which experience can be material to expert qualifications relates to the attrition of skills and knowledge over time. Mental health professionals often complete their training within several years of their thirtieth birthdays and may engage in practice, including the provision of expert testimony, over the subsequent four or five decades. Brief exposure to information about a particular disorder²³⁹ or some experience in evaluating and treating the condition may fade from memory several decades later without continued experience. Just as important is the possibility that additional knowledge about the condition has been gained in the interim and that additional knowledge will make an important contribution to the expert's evaluation. Training on a mental disorder or treatment, therefore, may be a necessary but insufficient aspect of an expert's qualifications in the absence of ongoing experience. Ongoing experience may include evaluating or treating patients with the disorder, teaching trainees how to assess or treat the disorder, systematically reviewing the literature on the disorder, attending continuing-education sessions about the disorder, conducting research on the disorder, and conducting forensic evaluations on individuals with the disorder.

But there is also a danger that experience can be overemphasized as a criterion of expertise. Assuming a baseline degree of adequate training and some ongoing experience in a field or with a condition, it is not clear that additional experience necessarily enhances an expert's authoritativeness. Experts will sometimes cite the number of evaluations they have performed of a particular type of evaluatee (e.g., alleged or convicted murderers) or of a given kind (e.g., assessments of competence to stand trial). However, if evaluations are performed inadequately or are unduly influenced by various cognitive biases,²⁴⁰ mere experience may

239. Although this discussion is framed in terms of a particular disorder, the condition at issue may not be a disorder in the formal sense. It may instead involve a symptom (e.g., auditory hallucinations), a mental state not linked to a specific disorder (e.g., dissociation), or a behavioral propensity (e.g., violent behavior). The argument in this section is generally applicable to all these categories of phenomena.

240. Cognitive biases in this context involve the systematic influence on human decision-makers under conditions of uncertainty, which describe the context in which forensic mental health assessments are conducted. Examples of such biases include confirmation bias (the tendency to selectively emphasize evidence consistent with an established position and de-emphasize evidence that refutes it), retention bias (favoring the side retaining the expert), and blind spot bias (the belief that cognitive biases do not apply to oneself in the same way they do to others). See, e.g., Itiel E. Dror, Saul M. Kassir & Jeff Kukucka, *New Application of Psychology to Law: Improving Forensic Evidence and Expert Witness Contributions*, 2 J. Applied Rsch. Memory & Cognition 78 (2013), <https://doi.org/10.1016/j.jarmac.2013.02.003>; Tess M. Neal & Stanley L. Brodsky, *Forensic Psychologists' Perceptions of Bias and Potential Correction Strategies in Forensic Mental Health Evaluations*, 22 Psych. Pub. Pol. & L. 58 (2016), <http://dx.doi.org/10.1037/law0000077>; Tess M. Neal et al., *A General Model of Cognitive Bias in Human Judgment and Systematic Review Specific to Forensic Mental Health*, 46 L. Hum. Behav. 99 (2022), <https://doi.org/10.1037/lhb0000482>. See generally, Tess M. Neal et al., *Psychological Assessments in Legal Contexts: Are Courts Keeping "Junk Science" Out of the Courtroom?* 20 Psych. Sci. Pub. Int. 135 (2020), <https://doi.org/10.1177/1529100619888860>.

only reinforce bad practice. Studies of the relationship between experience and the performance of mental health professionals in areas such as diagnosis²⁴¹ and bias management²⁴² have not demonstrated a relationship between superior performance and experience in either area. It may be that, despite having less clinical experience, recently trained clinicians are more familiar with the contemporary diagnostic framework and the research on cognitive bias and are less tempted to use their clinical experience as a substitute for generally accepted criteria (e.g., “I know schizophrenia when I see it, regardless of what the criteria say.”).

Licensure and Board Certification

Licensure

Possession of a valid professional license is usually a threshold requirement for an expert to be considered qualified in legal proceedings. Licensure of physicians (including psychiatrists) is governed by a licensure board in each state.²⁴³ Although criteria may differ, generally a physician who has graduated from an accredited American medical school, passed a sequence of tests designed to ensure adequate levels of knowledge and clinical judgment,²⁴⁴ and completed one or two years of residency training is eligible for full licensure.²⁴⁵ Prior to that, a temporary license allowing practice under supervision is usually issued. Graduates of medical schools that are not in the United States are usually subject to different standards, often being required to spend longer periods in residency training and undergoing individual review of qualifications. Once a physician attains licensure in a state, the process to acquire a license in another state varies. Some states grant additional licenses fairly easily; others, such as California, require that physicians take and pass a test of general medical knowledge if a certain period of time has passed (e.g., ten years in California) since the original sequence of testing was completed.²⁴⁶

241. Here, reliability is being used in its technical sense of agreement across more than one rater. For an example of the failure to find a consistent effect of previous experience, see, e.g., Sean H. Yutzy et al., *DSM-IV Field Trial: Testing a New Proposal for Somatization Disorder*, 152 Am. J. Psychiatry 97 (1995).

242. See Patricia A. Zapf et al., *Cognitive Bias in Forensic Mental Health Assessment: Evaluator Beliefs About Its Nature and Scope*, 24 Psych. Pub. Pol. & L. 1 (2018), <http://dx.doi.org/10.1037/law0000153>.

243. A summary of the requirements for medical licensure in each jurisdiction is available from the Federation of State Medical Boards at <https://perma.cc/8UG4-CQU8>.

244. See a description of the tests and the examination process at <https://perma.cc/J362-C26L>.

245. Federation of State Medical Boards, *supra* note 243.

246. Calif. Bus. & Pro. Code § 2184 (2021).

For clinical psychologists, standards for licensure differ somewhat by state, but generally after completion of an accredited Ph.D. program in the United States (including a one-year internship), they are required to complete one year of clinical work under the supervision of a licensed psychologist and pass a national licensure examination.²⁴⁷ Because the states do not restrict the practice of psychotherapy per se, but instead regulate the use of professional titles, an unlicensed psychologist can engage in many aspects of the clinical practice of psychology, including all forms of psychotherapy, but will not be able to use the title of psychologist. For psychologists who are seeking licensure in another jurisdiction, some states will grant reciprocity—that is, they will not engage in an independent process of reviewing the applicant’s credentials, relying instead on the review conducted by the initial licensure board.

Board Certification

Board certification represents qualifications beyond those required for licensure in either medicine or psychology. Although well-trained, competent psychiatrists may have reasons for not attaining board certification (e.g., examination anxiety that interferes with performance; a career centered on nonclinical research, for which clinical board certification is thought to be unnecessary), the tests are designed to be passed by a competent psychiatrist and do not require exceptional levels of clinical skill. Thus, in most cases, one can consider board certification as reflecting that a psychiatrist has attained adequate clinical competence for independent psychiatric practice. Whether a court chooses to admit testimony from a psychiatrist who has not been board certified may depend on the reasons a psychiatrist has not been certified and on the specific question(s) that the psychiatrist’s testimony will address.²⁴⁸

Professional psychology also has a board-certification process, administered by the American Board of Professional Psychology.²⁴⁹ Certification is only offered in psychology specialties, but these include such general clinical fields as

247. Details of requirements in each state can be found at the website of the Association of State and Provincial Psychology Boards at <http://www.asppb.net>.

248. For examples of the scope of judicial discretion on this issue, see, e.g., *Hall v. Quarterman*, 534 F.3d 365 (5th Cir. 2008) (finding that a state requirement that only a licensed expert may testify in a civil commitment hearing as to mental retardation did not extend to expert testimony on the same topic); *Oberlander v. Oberlander*, 460 N.W.2d 400 (1990) (reversing as abuse of discretion the trial court’s exclusion of expert testimony from a psychologist who was licensed in the neighboring state); *Williams v. Brown*, 244 F. Supp. 2d 965 (N.D. Ill. 2003) (finding that psychiatrists who were not board-certified child psychiatrists may nonetheless testify about the condition of juvenile plaintiffs).

249. A description of the process and eligibility requirements for the examination process can be found at <https://perma.cc/D2ES-YCRS>.

clinical, counseling, and group psychology. As in subspecialty certification in psychiatry, candidates are expected to demonstrate advanced competence in the specialty area, defined specifically for each specialty. Board certification is far less common among psychologists than psychiatrists, in part because board certification is an expected part of the credentialing process in psychiatry, but it is not in psychology.²⁵⁰ It is therefore less likely that certification will be applied as a minimum standard for expert testimony in psychology than in psychiatry or other areas of medicine.

Specialty credentialing can be obtained in forensic psychiatry (from the American Board of Psychiatry and Neurology²⁵¹) and forensic psychology (from the American Board of Professional Psychology²⁵²). This particular specialization is often the most relevant to legal proceedings, as candidates must demonstrate the requisite post-training experience, competence in clinical areas, and knowledge of relevant law that guides forensic evaluations.²⁵³ Other board certification specialty areas in psychology (e.g., clinical psychology, clinical neuropsychology, police and public safety) and psychiatry (e.g., adult psychiatry, addiction psychiatry, child and adolescent psychiatry) may also be applicable, depending on the nature of the legal questions.

Prior Relationship with the Subject of the Evaluation

Some attorneys, judges, and jurors may presume that a mental health professional who has had a treatment relationship with the person whose mental state is in question is better qualified to testify about aspects of that mental state than an evaluator who is meeting the person for the first time. The logic seems strong: A professional who has known the person for some period of time should be better able to offer conclusions about the person's diagnosis and treatment requirements and the impact of the person's mental state on function and behavior. So it may seem surprising that the ethics guidelines of both the American Academy of Psychiatry and the Law (the leading organization of forensic psychiatrists) and the American Psychological Association point to problems

250. Approximately 85% of psychiatrists become board certified in the eight years following completion of residency training. Dorthea Juul, James H. Scully Jr. & Stephen C. Scheiber, *Achieving Board Certification in Psychiatry: A Cohort Study*, 160 Am. J. Psychiatry 563 (2003), <https://doi.org/10.1176/appi.ajp.160.3.563>. In contrast, it was estimated that in 2000 only 3.5% of psychologists had achieved board certification. Frank M. Dattilio, *Board Certification in Psychology: Is It Really Necessary?* 33 Pro. Psych. Rsch. & Prac. 54 (2002), <https://doi.org/10.1037/0735-7028.33.1.54>.

251. See <https://perma.cc/VH4H-6GUW>.

252. See <https://perma.cc/D2ES-YCRS>.

253. See, e.g., Melton et al., *supra* note 1; Gutheil & Appelbaum, *supra* note 69.

inherent in such situations (the *Ethical Principles of Psychology and Code of Conduct*,²⁵⁴ broadly applicable to all psychology, and the *Specialty Guidelines for Forensic Psychology*,²⁵⁵ which applies to the practice of psychology in legal contexts).²⁵⁶ Although none of these guidelines state that giving testimony about current or former patients is clearly unethical, they each offer words of caution and discourage clinicians from playing both clinical and expert roles.²⁵⁷

The professional literature on this issue, and the ethics guidelines themselves, cite several reasons why having a treating professional perform the evaluation for legal purposes may not be prudent.²⁵⁸ First, offering testimony, even if it is supportive of the patient's legal claim, may interfere with the therapeutic relationship.

254. American Psychological Association, *Ethical Principles of Psychologists and Code of Conduct* (2017). See <https://perma.cc/YYP9-3NR6>.

255. APA Specialty Guidelines, *supra* note 71.

256. *Id.*

257. The forensic psychiatry guidelines explicitly discourage this practice:

Psychiatrists who take on a forensic role for patients they are treating may adversely affect the therapeutic relationship with them. Forensic evaluations usually require interviewing corroborative sources, exposing information to public scrutiny, or subjecting evaluatees and the treatment itself to potentially damaging cross-examination. The forensic evaluation and the credibility of the practitioner may also be undermined by conflicts inherent in the differing clinical and forensic roles. Treating psychiatrists should therefore generally avoid acting as an expert witness for their patients or performing evaluations of their patients for legal purposes.

Forensic Psychiatry Guidelines, *supra* note 70, at IV. In contrast, the *Specialty Guidelines for Forensic Psychology* could be seen as somewhat more permissive:

A multiple relationship occurs when a forensic practitioner is in a professional role with a person and, at the same time or at a subsequent time, is in a different role with the same person; is involved in a personal, fiscal, or other relationship with an adverse party; at the same time is in a relationship with a person closely associated with or related to the person with whom the forensic practitioner has the professional relationship; or offers or agrees to enter into another relationship in the future with the person or a person closely associated with or related to the person (EPPCC Standard 3.05). Forensic practitioners strive to recognize the potential conflicts of interest and threats to objectivity inherent in multiple relationships. Forensic practitioners are encouraged to recognize that some personal and professional relationships may interfere with their ability to practice in a competent and impartial manner and they seek to minimize any detrimental effects by avoiding involvement in such matters whenever feasible or limiting their assistance in a manner that is consistent with professional obligations.

APA Specialty Guidelines, *supra* note 71, at 4.02.

258. Larry H. Strasburger, Thomas G. Gutheil & Archie Brodsky, *On Wearing Two Hats: Role Conflict in Serving as Both Psychotherapist and Expert Witness*, 154 Am. J. Psychiatry 448 (1997), <https://doi.org/10.1176/ajp.154.4.448>; Ronald Schouten, *Pitfalls of Clinical Practice: The Treating Clinician as Expert Witness*, 1 Harv. Rev. Psychiatry 64 (1993), <https://doi.org/10.3109/10673229309017058>; Stuart Greenberg & Daniel Shuman, *Irreconcilable Conflict Between Therapeutic and Forensic Roles*, 28 Pro. Psych.: Rsch. & Prac. 50 (1997), <https://doi.org/10.1037/0735-7028.28.1.50>; Gutheil & Appelbaum, *supra* note 69, at 243–47.

Second, the underlying assumption about the desirability of having the clinician testify may be flawed. Although the clinician may have known the person as a patient, the clinical process may never have required the clinician to collect the type of information that would be relevant to the legal question. Even if that information were discussed, the treating clinician is less likely to have approached it with the degree of caution that a forensic evaluator would be likely to employ or to have attempted to verify the information through collateral sources. Even after agreeing to participate as an expert witness, a clinician may be unaware of the importance of assessing the veracity of the person's claim or may be afraid that doing so could lead to strains in the therapeutic relationship.

A third problem is that the clinician, having formed an alliance with the person as a patient perhaps over a considerable period of time, may feel a natural allegiance to that person and a desire, even if not conscious, to support the person's contentions in the case. Thus, evidence presented may be subtly distorted or consciously manipulated by clinicians who see their role as being a patient's advocate.

Fourth, there is an ethical problem when a clinician is subpoenaed to testify over the patient's objection. The preexisting therapeutic relationship was premised on the information that the patient revealed being used for treatment purposes. Being subpoenaed to testify over the patient's objection places a clinician whose testimony cannot support the person's legal claim in an extremely awkward position, compelled now to use that information potentially to the patient's detriment.²⁵⁹

In contrast, therefore, to what might seem like the logical assumption—that a treating clinician is best qualified to testify regarding the patient—there are multiple reasons not only to avoid relying on treating clinicians, but to discourage them from serving as expert witnesses.

How Was the Assessment Conducted?

The reliability and validity of an expert's opinion on mental health issues depends greatly on how the assessment that forms the basis for the conclusions was conducted.

259. Although all states have psychotherapist-patient and/or physician-patient testimonial privilege statutes that limit testimony by treating psychiatrists and psychologists (and often other mental health professionals) without the patient's consent, the exceptions in many of these statutes—including the so-called patient-litigant exception that is invoked when patients place their mental state at issue in a case—are sufficiently numerous that this situation cannot be ruled out. *Jaffee v. Redmond*, 518 U.S. 1 n.13 (1996); Christopher B. Mueller et al., § 5.35 Psychotherapist-Patient Privilege, GWU Legal Studies Research Paper No. 2018–68 (2018).

Was the Evaluatee Examined in Person?

Given the range of cases in which mental health experts provide testimony and the various questions they must respond to, there are situations in which experts provide evidence without having examined the person about whom they are testifying.²⁶⁰ These circumstances may arise when direct evaluation is impossible—for example, in contests over testamentary capacity, where often a claim regarding a testator’s capacity will be litigated only after the person is deceased. Other civil litigation that can raise questions about the state of mind of a deceased person includes contractual capacity, wrongful death, and medical malpractice claims.²⁶¹ Testimony about a person who cannot be evaluated directly is less likely in criminal cases, but highly contentious examples occurred in death penalty cases in Texas: Defendants have the right to decline evaluation by prosecution experts,²⁶² but those experts may testify on the basis of a hypothetical question that incorporates some of the facts of the defendants’ history and behavior.²⁶³

Conclusions about persons who have not been directly examined may be drawn on the basis of available records, including medical, mental health, police, educational, armed services, and other records; information from informants who have been or are in contact with the person, which the expert may derive from interviews, prior testimony, depositions, police reports, and other sources; and on some occasions, the expert’s observations of the person’s behavior, for example, in a prison or courtroom setting.²⁶⁴ Although it may be possible to draw valid conclusions on the basis of such data, these conclusions are generally

260. On some occasions, testimony will provide contextual information for the decision-maker—for example, how a person in a given situation or with a given disorder would usually respond, without being applied directly to a specific person. John Monahan & Laurens Walker, *Social Science in Law: Cases and Materials* (10th ed. 2022).

261. Farnsworth, *supra* note 9, § 3:11. For a case study of the use of postmortem analysis in the USS *Iowa* explosion investigation, see Charles Patrick Ewing & Joseph T. McCann, *Minds on Trial: Great Cases in Law and Psychology* 129–39 (2006); *Moon v. United States*, 512 F. Supp. 140 (D. Nev. 1981) (finding that hospital psychiatrists were negligent in diagnosing with schizophrenia a patient who later committed suicide); *Urbach v. United States*, 869 F.2d 829 (5th Cir. 1989) (finding no medical malpractice where a mental patient on furlough from a VA hospital was arrested and beaten to death in a Mexican prison). See also Norman Poythress et al., *APA’s Expert Panel in the Congressional Review of the USS Iowa Incident*, 48 Am. Psych. 8 (1993), <https://doi.org/10.1037/0003-066X.48.1.8>.

262. *Estelle v. Smith*, 451 U.S. 454 (1981).

263. *Barefoot v. Estelle*, 463 U.S. 880 (1983); *Satterwhite v. Texas*, 486 U.S. 249 (1988).

264. Kirk Heilbrun et al., *Third-Party Information in Forensic Assessment*, in 11 *Handbook of Psychology: Forensic Psychology* 69 (Alan M. Goldstein ed., 2003). Testimony offered in capital sentencing contexts without examining the defendant has been particularly controversial. See, e.g., *Bennett v. State*, 766 S.W.2d 227, 232 (Tex. Crim. App. 1989) (Teague, J., dissenting) (“[W]hen Dr. Grigson testifies at the punishment stage of a capital murder trial he appears to the average lay juror . . . to be the second coming of the Almighty. . . . Dr. Grigson is extremely good at

more limited and have a lesser degree of certainty than when a direct evaluation has taken place. The ethics statements of the major forensic psychiatry and forensic psychology organizations offer words of caution about such testimony,²⁶⁵ and there are several reasons why caution is warranted.

One can think of expert knowledge in mental health as comprising two components: knowing how to conduct an evaluation to obtain relevant information and knowing how to weigh that information to reach a conclusion.²⁶⁶ When an expert cannot carry out an examination of the person, the expert must rely on information accumulated by others, sometimes for other purposes. It is very unlikely that all the data that the expert would have wanted to obtain will be available in such circumstances. This is true even with data gathered by another mental health professional (e.g., in medical or mental health records), because that person may not have asked all the questions that the testifying expert would have asked, fully recorded the responses, and considered the possibility of exaggeration or minimization by the individual being evaluated. The intangible aspects of an evaluation, including the person's relatedness, emotions, and degree of cooperation, may be especially difficult to convey. Because many of the diagnostic categories require that other possibilities have been excluded first,²⁶⁷ the

persuading jurors to vote to answer the [future dangerousness] issue in the affirmative.”); *They Call Him Dr. Death*, Time, June 1, 1981; Rosenbaum, *supra* note 166.

265. In the *Ethics Guidelines for the Practice of Forensic Psychiatry*, the American Academy of Psychiatry and the Law notes:

For certain evaluations (such as record reviews for malpractice cases), a personal examination is not required. In all other forensic evaluations, if, after appropriate effort, it is not feasible to conduct a personal examination, an opinion may nonetheless be rendered on the basis of other information. Under these circumstances, it is the responsibility of psychiatrists to make earnest efforts to ensure that their statements, opinions and any reports or testimony based on those opinions, clearly state that there was no personal examination and note any resulting limitations to their opinions.

Forensic Psychiatry Guidelines, *supra* note 70, at IV.

The comparable guidelines for forensic psychology state,

Forensic practitioners recognize their obligations to only provide written or oral evidence about the psychological characteristics of particular individuals when they have sufficient information or data to form an adequate foundation for those opinions or to substantiate their findings (EPPCC Standard 9.01). Forensic practitioners seek to make reasonable efforts to obtain such information or data, and they document their efforts to obtain it. When it is not possible or feasible to examine individuals about whom they are offering an opinion, forensic practitioners strive to make clear the impact of such limitations on the reliability and validity of their professional products, opinions, or testimony.

APA Specialty Guidelines, *supra* note 71, at 9.03.

266. Paul S. Appelbaum, *Hypotheticals, Psychiatric Testimony, and the Death Sentence*, 12 Bull. Am. Acad. Psychiatry & L. 169 (1984); see also Am. Psychiatric Ass'n amicus brief in *Barefoot*, 463 U.S. at 880.

267. For example, DSM-5-TR criteria for major depressive disorder require both that the symptoms on which a diagnosis is based not be due to the direct physiological effects of a drug (licit

absence of pertinent negative information (e.g., the person does not misuse substances) can restrict the ability to make definitive diagnoses and draw other relevant conclusions. Moreover, these problems are compounded when the data available to the expert have been shaped by someone with an interest in the outcome of the case, as when an expert testifies in sole reliance on information in a hypothetical question that is constructed to support the interests of one party in the litigation.

Thus, the major professional organizations in forensic mental health agree that evidence based on sources other than direct evaluation of the person should be framed with due regard for its limitations and that experts should make those limitations clear in their reports or testimony. An expert witness's failure to do so may be unethical and should probably cast doubt on the credibility of the evidence presented.

Did the Evaluatee Cooperate with the Assessment?

Even when a direct evaluation has taken place, the evaluatee's degree of cooperativeness may affect the validity of the data.²⁶⁸ Civil plaintiffs and criminal defendants have obvious reasons to distrust experts who are examining them on behalf of adverse parties and may be less than forthcoming (or provide information that is inaccurate). But even when an evaluation is being conducted by an expert retained by the person's own attorney, the person's cooperativeness may be limited by the symptoms of the disorder. For example, a person who is experiencing paranoid delusions may be suspicious and fearful even of an expert with whom that person's own attorney encourages cooperation (and even of the attorney). It is therefore important for the expert to clarify, when presenting evidence and conclusions based on the evaluation, the extent to which the evaluatee cooperated with the examination process. If the examinee was less than cooperative, it is also useful for the evaluator to explain why that might have been, and to describe the impact on the information gathered and the steps taken (often relying on other sources of information) to compensate for this lack of cooperation.

or illicit) that has been ingested or to a general medical condition; that they not be better accounted for by bereavement after the death of a loved one; and that the person has never experienced a manic or hypomanic episode. DSM-5-TR, *supra* note 72, at 183. Other major diagnostic categories carry similar requirements to rule out the possibility that the person's presentation is due to other causes before making the diagnosis in question.

268. Melton et al., *supra* note 1.

Was the Evaluation Conducted in Adequate Circumstances?

Mental health evaluations often involve discussions of sensitive material, including histories of abuse, use of illegal substances, sexual practices, intimate fears and fantasies, and potentially embarrassing symptoms. Some evaluations may also involve discussion of alleged or acknowledged criminal conduct. Although some persons may be reluctant to speak freely about these issues with an evaluator they barely know—and who may reveal this information in the courtroom—the reassurance that they are talking with a mental health professional often substantially mitigates those concerns.²⁶⁹ However, an evaluation conducted in a setting that is less than private lowers the likelihood of examinee disclosures.²⁷⁰ This is often a problem in correctional institutions, where interviews may be conducted in a place where correctional officers or other inmates can overhear them. Medical hospitals are another location where privacy may be compromised, with nursing staff or other patients nearby. Even if no one is within earshot, interview sites that are noisy or subject to other distractions may interfere with the evaluatee's ability to attend to the questions and respond accurately; this can be a particular problem for people with mental disorders that may impair concentration and attention. Whenever possible, a competent evaluator tries to obtain a venue that is free of these intrusions, and when it is not possible, the situation should be noted as a limitation on the completeness of the evaluation in the report or testimony. This challenge was exacerbated during the Covid-19 era, during which time remotely conducted evaluations became more frequent. During remote evaluations, an examiner can be less certain about who is within hearing range and may need to rely on the account of the individual examined as well as any available staff member regarding whether there is privacy.

Attorneys sometimes ask to sit in on an evaluation. Their presence can raise similar concerns, even when they are representing the person being evaluated, because the type of information discussed in a mental health evaluation may be quite different from what a client usually discloses to an attorney.²⁷¹ Particularly

269. There is some research on the question of whether evaluatees may too easily be induced to speak frankly with someone who is introduced as a mental health professional, but whose role is very different than would obtain in treatment settings and who may reach opinions adverse to the person's interests. See, e.g., Daniel Shuman, *The Use of Empathy in Forensic Evaluations*, 3 *Ethics & Behav.* 289 (1993), <https://doi.org/10.1080/10508422.1993.9652109>; Strasburger et al., *supra* note 258; Greenberg & Shuman, *supra* note 258.

270. Melton et al., *supra* note 1. Distraction can be a particular problem when formal psychological tests are used; see, e.g., Kirk Heilbrun, *The Role of Psychological Testing in Forensic Assessment*, 16 *L. & Hum. Behav.* 257 (1992), <https://doi.org/10.1007/BF01044769>.

271. Robert I. Simon, "Three's a Crowd": *The Presence of Third Parties During the Forensic Psychiatric Examination*, 24 *J. Psychiatry & L.* 3 (1996), <https://doi.org/10.1177/009318539602400102>.

when the examination is being conducted by an expert for an adverse party, attorneys may be tempted to object to questions or signal their clients regarding their answers. To help mitigate this, if an attorney is present, as is sometimes unavoidable, the ground rules should include that the attorney sits out of the line of sight of the evaluatee and does not interrupt the examination. An alternative is to have the evaluation audiotaped or videotaped, a technique that some experts routinely use. There is a lack of data on the impact of taping on evaluatees' willingness to be forthcoming, but experienced forensic examiners have expressed the view that evaluatees rapidly adjust to the recording equipment, with little impact on the evaluation.²⁷²

A final consideration is the amount of time available for the examination.²⁷³ Time constraints may result from correctional rules (e.g., prisoners are available only during given periods of time; the availability of equipment used to conduct a remote evaluation is limited relative to the demand), medical illnesses or mental disorders (e.g., the evaluatee has limited strength or attention), or limitations on resources (e.g., the party employing the expert has funds only for a certain number of hours of work). Appropriate duration of an examination is likely to depend on the questions being asked, the complexity of the person's history and presentation, and the person's cooperation with the evaluation. The duration of an examination, standing alone, is not a good indicator either of its quality or of the validity of the conclusions that were drawn. But an expert should be able to assess the time necessary to perform an adequate evaluation and, if sufficient time is not available, should indicate the limitations on the resulting opinions.

Were the Appropriate Records Reviewed?

The importance of the evaluator having access to the person's records will vary somewhat depending on the legal question being addressed, but it can often be critical to the validity of the evaluation.²⁷⁴ When retrospective assessments are being conducted—for example, an evaluation of a defendant's state of mind at the time of a crime that occurred months to years before the examination, or an assessment of a person's capacity to enter into a contract at some distant prior date—reviewing contemporary or nearly contemporary records can provide crucial insights into the person's symptoms and functioning at that time. But even when contemporaneous function or future behavior is being assessed,

272. AAPL Task Force, *Videotaping of Forensic Psychiatric Evaluations*, 27 J. Am. Acad. Psychiatry & L. 345 (1999).

273. Melton et al., *supra* note 1, at 47.

274. Heilbrun et al., *supra* note 264; see also discussion in section titled "Data Collection and Sources of Information" above.

having access to available records may still be of great importance. Because distinctions between mental disorders can depend in part on the pattern of symptoms over time, accurate diagnosis is often dependent on having a view of the person's prior psychiatric history.²⁷⁵ In addition, when malingering or defensiveness are active considerations, as they often are, the consistency of the person's presentation over time can be an important factor in the assessment.²⁷⁶ And given that past behavior is generally the best predictor of future behavior, especially where violence is concerned, knowledge of a person's previous history can be essential for predictions of reasonable accuracy.²⁷⁷ Thus, regardless of the focus of the evaluation, an effort should be made to obtain all relevant available records.

Which records are relevant will depend somewhat on the legal question being asked.²⁷⁸ Whenever possible, records of past mental health evaluations or treatment should be obtained. Medical records often contain information about psychiatric symptoms, alcohol and drug use, and functional levels, and thus can be useful. Educational, work, and military records can usefully inform both patterns of symptoms and functional impairment. Educational records may be especially helpful where disorders of early onset are suspected, and work and military records are often useful when occupational disability is at issue. In criminal cases, particularly those involving assessments of the defendant's state of mind at the time of the crime, police records can often be valuable, including interviews with witnesses or the defendant, and the results of physical evaluations—including pictures—of the crime scene. It can be helpful to compare the data obtained by these means with the defendant's accounts of the episode that led to the arrest. Diaries or other accounts written by the person whose mental state is at issue are sometimes available and, to the extent they were generated prior to the initiation of legal proceedings, can be enlightening regarding the person's state of mind and motivation, the influence of third parties, and the like. When there has been prior litigation involving the person being evaluated, depositions or transcripts of testimony can be helpful for information about state of mind and factual data. Records of prior incarceration are useful in appraising likely adjustment and possibly response to rehabilitative interventions if the evaluatee is again incarcerated.

275. Diagnosis and subcategorization of bipolar disorder, for example, are dependent not only on assessing the person's current symptoms—whether manic or depressed—but also on ascertaining whether mania or depression was present in the past if it is not apparent at present. See DSM-5-TR, *supra* note 72, at 139–43.

276. See generally section titled “Response Style” above.

277. See generally section titled “Predictive Assessments” above.

278. See, e.g., Melton et al., *supra* note 1.

Was Information Gathered from Collateral Informants?

In addition to reviewing records, interviewing informants with relevant data can provide important perspectives on the person being evaluated.²⁷⁹ Family members, friends, and coworkers often can report on patterns of behavior indicative of symptoms of mental disorder or functional impairment. They may know about prior behavioral health treatment or histories of involvement with the criminal justice system. Current or former therapists can share useful impressions of diagnosis and comment on levels of function and response to treatment, although to the extent that their interactions with the person are subsumed under a psychotherapist-patient or physician-patient privilege, and do not fall under one of the exceptions in that jurisdiction, it may not be possible to contact them without the person's consent. Witnesses to an alleged crime or workplace harassment can similarly round out a picture of the person and help to confirm or disconfirm the evaluator's impressions. Access to collateral informants may be complicated by legal restrictions or, if they are close to the person being evaluated, by their reluctance to speak to an expert working for an adverse party. When contact does occur, the evaluator needs to consider possible distortions by the informant in the service of helping, or sometimes of harming, the interests of the person who is the subject of the evaluation. These challenges notwithstanding, collateral interviews can provide valuable historical and current information that is often difficult or impossible to obtain from records.

Were Medical Diagnostic Tests Performed?

Dualistic views of human behavior, in which mind and body are seen as distinctly separate entities, have been rejected by scientists who study thought and behavior and clinicians who treat behavioral health disorders.²⁸⁰ The relevant fields, including cognitive science, neuroscience, psychology, psychiatry, and philosophy, now acknowledge the brain as the seat of mentation and behavior and recognize that all mental phenomena, including abnormal mental states, result from perturbations in the function of the brain. At some level, there must be a physical concomitant of every mental phenomenon, and sometimes the physical influences on abnormal behavior are gross enough to be detected by existing techniques, which may reveal potentially treatable conditions. Thus, to identify the causes of abnormal thought or behavior and formulate a diagnosis

279. Heilbrun et al., *supra* note 264.

280. See, e.g., Kenneth S. Kendler, *Toward a Philosophical Structure for Psychiatry*, 162 Am. J. Psychiatry 433 (2005), <https://doi.org/10.1176/appi.ajp.162.3.433>.

may require an evaluation of a person's physical state, along with the mental state.²⁸¹ If there is any reason to suspect that an identifiable general medical disorder lies at the root of a person's condition (e.g., a sudden and unprecedented appearance of symptoms, disproportionate impairment of aspects of cognitive function), medical testing, including electroencephalograms (EEGs) and imaging studies, may be indicated.²⁸² Any medical diagnostic testing should be justified based on its relevance to the evaluation. Persons without medical training are typically not able to order or interpret the results of medical tests, which will require referral to a health professional if indicated.

Was the Evaluatee's Functional Impairment Assessed Directly?

As previously discussed, mental health evidence will often focus on the extent to which a person is capable of performing a particular task or set of tasks, and a person's impairment on one or more functional abilities.²⁸³ Sometimes an evaluator will be able to infer from an examination of the person's mental state and information from other sources whether the person is or was capable of performing the task at hand (e.g., standing trial, returning to work, managing property). However, direct assessment of the relevant function is another option for evaluation.²⁸⁴ Where a functional ability is at issue that relates to a discrete task or set of tasks, a competent evaluator should have considered using direct assessment of the person's performance of those tasks and should be able to explain any decision not to do so. It should be noted, though, that conclusions drawn even from direct assessments of function require some inference. A person claiming occupational impairment as a result of anxiety induced by long-standing harassment on the job, for example, might respond very differently to the demands of a

281. See generally section titled "Diagnosis of Mental Disorders" above.

282. Identification of structural or electrical abnormalities, however, does not necessarily imply that these things impaired the person's functioning or were responsible for the person's behavior. For discussion of a well-known case in which this issue was raised, see Stephen Morse, *Brain and Blame*, 84 Geo. L. J. 527 (1996). For a more general discussion of the introduction of findings of abnormalities demonstrated on brain imaging in court, see Dean Mobbs, *Law, Responsibility and the Brain*, 5 PLoS Biology 693 (2007), <https://doi.org/10.1371/journal.pbio.0050103>. As with structural findings, the mere presence of a functional abnormality is insufficient to establish a causal link to the person's mentation or behavior. There are growing legal and neuroscience literatures on the use of functional imaging data in court. See, e.g., Darby Aono, Gideon Yaffe & Hedy Kober, *Neuroscientific Evidence in the Courtroom: A Review*, 4 Cognitive Rsch. 1 (2019), <https://doi.org/10.1186/s41235-019-0179-y>; Jane Campbell Moriarty, *Neuroimaging Evidence in U.S. Courts*, in *Law and Mind: A Survey of Law and the Cognitive Sciences* (Bartosz Brozek et al. eds., 2021).

283. See generally section titled "Functional Impairment Due to Mental Disorders" above.

284. See section titled "Assessment of Functional Legal Capacities" above.

work-related task in the actual workplace compared to the safe confines of a mental health professional's office. Therefore, when employing direct assessment of functional capacity, the evaluator should be prepared to comment on the ecological validity of the test, that is, the degree to which the testing environment resembled the real-world environment in the person's life.²⁸⁵ Although observations in very different settings may have some value as part of the broader data set available in an evaluation, they do not carry the same weight as conclusions reached in environments similar to those at issue in the case.

Was the Possibility of Malingering or Defensiveness Considered?

In almost every mental health evaluation for legal purposes, the person being evaluated has an incentive to exaggerate or fabricate symptoms, to minimize or deny symptoms, or to distort the impact of actual symptoms on the person's functional abilities.²⁸⁶ The evaluator should consider the possibility of malingering or denial in every assessment, depending on the nature of the incentive. Techniques for appraising these different response styles were described earlier.²⁸⁷ Although these techniques are not foolproof, and well-prepared evaluatees can sometimes mislead mental health professionals on the existence or severity of disorders, successful malingering or minimizing over time are difficult tasks. Uncovering distortions of the degree of actual symptoms or of their impact is usually more challenging than detecting wholesale invention of disorders that are not present or adamant denials of disorders that are. Competent evaluators should be able to explain how they considered the possibility of distorted responding and why they believe that their conclusions are valid, while acknowledging that their degree of certainty can never be absolute.

285. Additional issues related to the use of functional tests are discussed below in the section titled "Was a Structured Diagnostic or Functional Assessment Instrument or Test Used?"

286. There are particular situations in which the incentive runs toward defensiveness rather than malingering. For example, a defendant facing relatively minor charges for whom an evaluation of competence to stand trial was ordered may have every reason to minimize the level of symptoms, preferring to go to trial quickly rather than spend an extended period of time in a psychiatric facility being treated to restore competence. A second example is a defendant whose risk for violence is being evaluated prior to a bail hearing, who also has a powerful incentive to downplay the presence of risk factors associated with violence and to minimize a past history of violence.

287. See section titled "Response Style" above.

Was a Structured Diagnostic or Functional Assessment Instrument or Test Used?

Notwithstanding the advantages of structured assessment techniques, they raise a set of concerns that must be addressed to determine their relevance to the question at issue and the weight that should be given to their results.

Has the Reliability and Validity of the Instrument or Test Been Established?

Reliability and validity are key concepts in test development.²⁸⁸ Each contains several subcategories. Reliability refers to the reproducibility of results obtained with a particular test. In other words, it is an estimate of the precision of an assessment technique. Interrater reliability is a measure of whether different examiners using the same test or instrument with the same subject obtain similar results, an important characteristic for an assessment approach that will be used by many raters. Test–retest reliability assesses the stability of results from an instrument or test over time; poor correspondence of results between time periods may indicate either an unreliable technique or a condition subject to periodic changes in status. It is an axiom of test and instrument development that good reliability is a prerequisite for having a valid assessment technique, but does not in itself guarantee validity.

Validity connotes the degree to which an instrument or test yields results that accurately reflect reality. Construct validity refers to the extent that an instrument or test reflects the theoretical construct that it purports to measure (e.g., anxiety or depression). Elements of construct validity include discriminant validity, which is the degree to which the test distinguishes between related conditions or states, and convergent validity, the extent to which the results of this test resemble results of other instruments that assess the same or a similar construct. Content validity describes the adequacy or thoroughness with which a test has sampled the variables associated with a given domain (e.g., does a measure of ability to work assess all relevant aspects of a given occupation?). Finally, predictive validity denotes the ability of an instrument or test to foretell a person's condition or behavior at some point in the future.

When the results of an evaluation using an instrument or test are offered in evidence, clarifying the extent to which reliability and validity have been

288. For the discussion in the following two paragraphs, see generally American Psychological Association, *Standards for Educational and Psychological Testing* (2014).

demonstrated is an essential aspect of determining admissibility and weight. In *Daubert*, when the U.S. Supreme Court referred to the reliability of a scientific technique, it was encompassing both reliability and validity as usually understood in the behavioral sciences.²⁸⁹ Which aspects of reliability and validity are relevant to a particular case will depend on the purpose for which the data from the test are being introduced. For example, if the evidence is addressing change in a person's test results over time, a measure's test-retest reliability becomes crucial. If more than one evaluator was involved, interrater reliability may be key. Discriminant validity will be relevant when two states or conditions must be distinguished from each other and predictive validity when forecasts of future mental state or behavior are being made. Careful evaluators will use only instruments or tests that have had the relevant types of reliability and validity confirmed in peer-reviewed publications, and they will be prepared to cite such data should questions be raised. Of course, some tests are so widely used over a sustained period that their reliability and validity are generally accepted (e.g., the MMPI-3). However, the reliability and validity of some long-standing tests (e.g., the Rorschach test) remain controversial,²⁹⁰ and data even from established tests can be used to reach conclusions of uncertain validity. Thus, novel uses of instruments or tests may also require demonstration of their psychometric characteristics for that purpose.

Functional assessment instruments are validated using outcomes involving the capacity to perform in a way that is directly relevant to the legal demands.²⁹¹ The reliability and validity of such a measure should be described in a manual that is either commercially available (along with the measure itself) to qualified professional purchasers, or available on a website or other publicly available source that describes the research establishing its reliability and validity.

Does the Person Being Evaluated Resemble the Population for Which the Instrument or Test Was Developed?

Even when present for one population, reliability and validity do not necessarily apply to those in a different population. If an assessment technique is being used

289. *Daubert v. Merrell Dow Pharms., Inc.*, 509 U.S. 579, 589 (1993).

290. Scott O. Lilienfeld, James M. Wood & Howard N. Garb, *The Scientific Status of Projective Techniques*, 1 Psych. Sci. Pub. Int. 27 (2000), <https://doi.org/10.1111/1529-1006.002>.

291. For example, a functional assessment instrument for competence to stand trial would measure capacities such as the ability to understand legal charges and how the legal system functions, assist counsel through communication and appropriate behavior during a legal proceeding, and make decisions regarding the options for plea. See, e.g., Grisso, *supra* note 3.

on someone from a different population than the one for which the instrument or test was developed, and the new group is likely to differ in some material way, reliability or validity may need to be reestablished. An example with regard to reliability might be using an instrument with a child that was developed to measure symptoms of mental disorders in adults.²⁹² Either the nature of the symptoms that adults experience or the ability of adults to describe their symptoms could be substantially different for children. Thus, it might be prudent for an evaluator to ascertain that data exist showing good reliability in this new population before using the assessment approach. An example involving validity is using predictive scales, such as instruments to assess violence risk, with a different group than the one on which the predictive algorithm was derived.²⁹³ Concretely, if a predictive test is based on a group of justice-involved individuals without serious mental illness, applying it to persons with serious mental illness—for whom different variables may affect behavior—is dubious in the absence of data demonstrating its validity in the latter group.

It should be emphasized, however, that reestablishing reliability and validity is necessary only when the original group and the new population are likely to differ in a relevant way. Why an instrument developed in California, for example, would not be as reliable and valid when used in Texas is not at all clear. The nature of the instrument or test also plays a role. Diagnostic tests are likely to differ in their characteristics across populations only if the disorders or the ways in which they manifest are different, which will not usually be the case. Predictive tests, however, may be more sensitive to cultural, socioeconomic, geographic, and other considerations that could introduce other influences on future conditions or behaviors. In addition, tests that involve comparisons with broader populations are said to be normed using those groups,²⁹⁴ and the comparative data (e.g., the evaluatee is in the lowest quartile of performance) may be invalid unless the test is renormed for the group of which the person is a member. The use of such tests in a clinical context, such as diagnosis for treatment-planning purposes, may differ from their use in legal settings. Thus, whether additional reliability and validity testing is required for a new use, or whether a test must be renormed before being used in this way, is necessarily a fact-specific determination.

292. The frequently differing presentations of mental disorders in children have led to the development of instruments intended specifically for use in that population. See, e.g., David Shaffer et al., *NIMH Diagnostic Interview Schedule for Children, Version IV (NIMH DISC-IV): Description, Differences from Previous Versions, and Reliability of Some Common Diagnoses*, 39 *J. Am. Acad. Child & Adolescent Psychiatry* 28, 28–38 (2000), <https://doi.org/10.1097/00004583-200001000-00014>.

293. See, e.g., John Monahan et al., *The Classification of Violence Risk*, 24 *Behav. Sci. & L.* 721 (2006), <https://doi.org/10.1002/bsl.725>.

294. For a good discussion of norming in the forensic context, see Grisso, *supra* note 3, at 56–59.

Was the Instrument or Test Used as Intended by Its Developers?

Established reliability and validity are necessary but not sufficient to determine whether an instrument or test has yielded reliable and valid results. Unless the assessment approach was applied in the manner intended by the developers, the data on reliability and validity may simply not be applicable to a particular use. Three possible areas of deviation relate to training in, administration of, and scoring of the assessment tool.

Training

Some instruments and tests are straightforward, requiring little specialized training. Familiarity with the manual accompanying the assessment tool might be sufficient. With other measures, though, training is required to ask the questions properly, especially when follow-up probing of responses is necessary or when evaluatees are asked to perform tasks that must be conducted in a particular way. Diagnostic instruments, in particular, may have complex “skip-out” rules, that is, procedures for determining when to include or omit certain questions based on the person’s responses to previous questions.²⁹⁵ When information is acquired at least in part from existing records rather than from the evaluatee directly, rules may exist for how the information should be identified and abstracted. These characteristics of an assessment approach may require training for proper administration, scoring, and interpretation.²⁹⁶ For more complex instruments or tests, assessors need face-to-face training with the opportunity to practice administration and receive supervision. Developers of such instruments or tests may offer training seminars for professionals.²⁹⁷ So a key question in assessing data based on an instrument or test is whether one requires special training to use it, and if so, whether the assessor was trained in the technique.

295. Structured diagnostic interviews have been widely used in epidemiological studies of mental disorders in the United States, as an example. See a description of a recent iteration of such a structured interview (the Structured Clinical Interview for DSM-5) at <https://www.appi.org/products/structured-clinical-interview-for-dsm-5-scid-5>.

296. Indeed, some psychological and neuropsychological tests should be administered only by psychologists trained in their use.

297. The creator of the popular Psychopathy Check List (PCL-R), for example, offers an extensive training program for clinicians and researchers desiring to learn how to properly administer the instrument. See <https://perma.cc/WSP8-JJJE>.

Administration

Even if the assessor was trained, the reliability and validity of an instrument or test will depend on whether it was administered properly. Many assessment tools require that questions be asked in a given sequence and phrased in a particular way. After an incorrect response, it may be permissible to ask the question again, but only a certain number of times. Probing of responses may be needed, but only certain probes may be permitted. Some tests are timed, with a given period allotted for the completion of a task. Deviations from any of these requirements could make the published data on the psychometric characteristics of the tool inapplicable to its use in a particular instance. Thus, a second crucial question is whether the instrument or test was administered in the same way as it was when its reliability and validity were established.

There are two ways this applies to the administration of tests used in forensic evaluations. First, these evaluations are often conducted in secure settings: jails, prisons, and detention centers. The evaluator must ensure that administration conditions are reasonably quiet, private, and distraction-free so that they do not deviate substantially from the conditions under which the test was validated. Second, the Covid-19 era witnessed an increase in the use of videoconferencing to conduct evaluations. When using such an approach, the evaluator must first determine whether conditions (as just discussed) *are* private, quiet, and distraction-free.²⁹⁸ Some tests for which the test-taker reads the items and responds on an answer sheet should not differ meaningfully whether administered in person or remotely; in fact, some such tests were administered via computer prior to Covid-19. Other tests cannot be translated as readily, and evaluators must consider the relevant research and the nature of the assessment to determine whether and to what extent such tests can be used during a remote evaluation.²⁹⁹

298. It is not unusual, for example, for the staff of a jail or prison to indicate to an evaluator that facility policy requires the presence of a correctional officer in the room. This is highly problematic, of course, as certain aspects of forensic evaluations involve very sensitive material. If this cannot be renegotiated with the facility, then the evaluator must make the difficult decision between conducting a modified evaluation under these circumstances or declining entirely.

299. Publishers of commercially available psychological tests (to qualified users) have worked since the onset of Covid-19 conditions in March 2020 to improve the technology that would address influences that adversely affect the conditions of remote administration, and hence make reliability and validity more comparable to in-person administration. *See, e.g.,* <https://perma.cc/6DLV-QV8F>. Yet there have not been innovations that would allow remote administration of tests requiring the observation of physical movements, hand-eye coordination, and the like—meaning that many neuropsychological tests cannot be administered remotely. *See, e.g.,* Kirk Heilbrun et al., *A Principles-Based Analysis of Change in Forensic Mental Health Assessment During a Global Pandemic*, 48 J. Am. Acad. Psychiatry & L. 293 (2020), <https://doi.org/10.29158/JAAPL.200039-20>.

Scoring

Assessment tools generally require that evaluatees' responses be scored in some way. For some instruments and tests, the scoring is simple and self-evident—for example, the number of positive responses is totaled to yield the score, or evaluatees themselves are asked to indicate the severity of their symptoms on a one-to-seven scale. Frequently the results are then calculated using computer software that automatically applies the relevant algorithm, generates statistical data, and even draws comparisons with broader groups, such as the general population or persons with a particular disorder. But there may be more complex scoring rules, particularly when evaluatees' verbal or narrative responses are elicited. An instrument assessing the severity of symptoms, for example, may require the person administering it to categorize responses along a numerical scale,³⁰⁰ and specific capacity-assessment tools frequently require similar judgments to be applied.³⁰¹ Published data on the reliability of scoring apply when the person administering the instrument adheres to the usual rules, and may not apply without such adherence. Without it, reliability may be adversely affected, and the results may be invalid as well. Hence, a third important question when such evidence is introduced is whether the rules for scoring responses were properly applied.

How Was the Expert's Judgment Reached Regarding the Legally Relevant Question?

As noted in the preceding sections, mental health experts' training and manner of conducting assessments is vital information when evaluating their testimony. However, the value of an expert's opinion also depends on the process by which the data were assessed and a conclusion was reached.

300. E.g., the Brief Psychiatric Rating Scale. See John E. Overall & Donald R. Gorham, *The Brief Psychiatric Rating Scale (BPRS): Recent Developments in Ascertainment and Scaling*, 24 *Psychopharmacology Bull.* 97 (1988), <https://doi.org/10.2466/pr0.1962.10.3.799>.

301. E.g., Grisso & Appelbaum, *supra* note 122.

Were the Findings of the Assessment Applied Appropriately to the Question?

Were Diagnostic and Functional Issues Distinguished?

Mental health professionals without experience in performing particular forensic evaluations may fail to recognize that the legal question being asked deals with a person's functional capacity, not with some aspect of their clinical state *per se*.³⁰² As a result, they may mistakenly base their opinions on the presence of a particular diagnosis or symptom cluster rather than on the person's capacity to perform in the legally relevant manner. Studies indicate that this has happened frequently in testimony on defendants' competence to stand trial, with experts often concluding that any psychotic defendant was *ipso facto* incapable of proceeding to trial.³⁰³ Similar problems may occur in hearings on guardianship or contests regarding testimonial capacity, where a person's ability to manage or dispose of assets might be thought incorrectly to turn solely on whether a neurocognitive disorder is present, as opposed to whether the person retains the necessary capacities despite the condition.³⁰⁴ This may be more likely to occur—and to go undetected—when experts are allowed or encouraged to address the ultimate legal issue in their testimony.³⁰⁵ When experts are permitted to testify to the ultimate question, the importance of probing their reasoning is magnified.³⁰⁶ Experts can be asked to identify the relevant functional capacities and to speak directly to the impact of the person's mental state on those capacities.³⁰⁷ This allows their reasoning processes and the correctness of their assumptions about the relevant functional standard to be tested.

302. See *Dusky v. United States*, 362 U.S. 402 (1960); Melton et al., *supra* note 1.

303. See, e.g., A. Louis McGarry, *Competency for Trial and Due Process Via the State Hospital*, 122 Am. J. Psychiatry 623 (1965), <https://doi.org/10.1176/ajp.122.6.623>. Subsequent studies suggested that this became a less common problem as educational efforts among mental health professionals who do such work had a positive impact. Robert A. Nicholson & Karen E. Kugler, *Competent and Incompetent Criminal Defendants: A Quantitative Review of Comparative Research*, 109 Psych. Bull. 355 (1991), <https://doi.org/10.1037/0033-2909.109.3.355>.

304. See Parry & Drogin, *supra* note 9, at 149–51.

305. See section titled “Ultimate-Issue Testimony” above.

306. See Parry & Drogin, *supra* note 9, at 429–31.

307. Buchanan, *supra* note 216.

Were the Limitations of the Assessment and the Conclusions Acknowledged?

Assessments are imperfect. Evaluatees may be less than cooperative. Records are often unavailable. Evidence from witnesses is conflicting. Time available is inadequate. Or the evaluator may simply have forgotten to ask about some piece of information that would have been helpful. Experts should be able to identify the limitations of their evaluations and the possible impact of those less-than-optimal aspects of the assessments. An expert would be unlikely to offer testimony if the expert believed that the limitations rendered the opinions invalid. But competent experts should be able to explain why, despite the limitations (which occur even in the best evaluations by the most capable experts), their evaluations were adequate to allow them to draw the conclusions that they intend to present.

A comparable set of limitations can occur when conclusions are drawn and opinions formulated. Just as all assessment tools have error rates, so do expert witnesses, although their rates are difficult to subject to statistical analysis. Errors may be introduced by inadequacies in the data available or the uncertainties inherent in particular determinations, especially predictions of future mental states and behaviors. As noted earlier, it is often impossible to specify the contingencies that may arise in a person's life that could influence their mental states and actions. Thus, any prediction—no matter how firmly grounded in available data—has a degree of uncertainty attached to it that a competent expert should be expected to acknowledge.

Are Opinions Based on Valid Empirical Data Rather Than Theoretical Formulations?

From the development of Freud's theories in the late nineteenth and early twentieth centuries until the present, many mental health professionals have based their clinical approaches on psychoanalytically inspired concepts. Some of these concepts have been confirmed scientifically (e.g., the existence of unconscious mental states), whereas others have not (e.g., dreams always represent the fantasied fulfillment of wishes). Although psychoanalytical theories and the psychodynamic psychotherapies that derive from them have declined in popularity in recent decades, many mental health professionals have received psychodynamic training and use these concepts to assess and treat their patients. Regardless of the possible clinical utility of these theories, which is controversial and may depend on the condition being treated, they are arguably more problematic when they serve as the basis for conclusions offered as part of legal proceedings. Nor are psychoanalytical theories the only ones that mental health professionals use and that may have a greater or lesser degree of empirical support.

To the extent that expert opinions are introduced to inform the judgments of legal factfinders, it is important for them to be based insofar as possible on

empirically validated procedures rather than on untested or untestable theories. That appears to be the import of the U.S. Supreme Court's decision in *Kumho Tire*.³⁰⁸ As Slobogin plausibly maintains, some legal questions (such as those concerning past mental states) may not easily lend themselves to approaches based on scientific methods, but expert opinions may nonetheless be of assistance to factfinders.³⁰⁹ At a minimum, it would seem fair for an expert to indicate when that is the case, so that the factfinder can make an informed judgment about the appropriate degree of reliance on the expert's opinion. And when empirically tested approaches are available, it is good practice for an expert to use them, or to explain why they were not used.

Case Example

Facts of the Case

James Stallworth³¹⁰ is a forty-eight-year-old male who was charged with making false statements and with fraudulent use of an identity document, incurring federal charges to which he pled guilty. His attorney requested an evaluation to possibly support the defense's motion for a downward departure at sentencing. The evaluation requested by the defense was conducted by a Ph.D. psychologist board certified in clinical psychology and forensic psychology (Dr. A). The assistant U.S. attorney requested a second evaluation, which was conducted by a psychiatrist in private practice who was board certified in psychiatry and forensic psychiatry and whose work includes both forensic evaluations and psychiatric treatment (Dr. B).

Mr. Stallworth does not have a juvenile record, but he has a substantial arrest history as an adult, resulting in fifteen convictions for offenses such as trafficking cocaine, resisting arrest, trespassing, possession of marijuana, burglary, theft, making false report, disorderly conduct resisting arrest, DUI, and shoplifting. He is also a registered Step 2 sex offender, having been sentenced to three years in prison for a sexual offense.

He presented for evaluation with Dr. A., who interviewed him on two occasions; Mr. Stallworth was polite and cooperative. His speech was somewhat rambling during the first interview; it was very rambling and tangential during the second. He was largely calm throughout, displaying no difficulty with attention or concentration, and worked consistently on all tasks. Intellectual

308. *Kumho Tire Co. v. Carmichael*, 526 U.S. 137 (1999) (holding that the *Daubert* standard for admitting expert testimony also applies to nonscientists).

309. Slobogin, *supra* note 212.

310. See Kirk Heilbrun et al., *Forensic Mental Health Assessment: A Casebook* 100–109 (2d ed. 2014). This is a composite of several cases and has been fully deidentified. There is no “James Stallworth.”

functioning was not formally measured but appeared to be in the Average to High Average range, as suggested by his appearance, behavior, and academic skills measured by the Wide Range Achievement Test—4th edition (Word Reading, 13th grade equivalent; Sentence Comprehension, 13th grade equivalent; Math, 13th grade equivalent; Spelling, 11.6th grade equivalent). He did not report experiencing perceptual disturbances at the time of the evaluation, and there was no evidence of bizarre ideas, delusions, or other thought disturbances.

He described attending high school and a total of about three years of college at the University of South Florida and the University of Delaware. He said that he had never been married but was described by a friend as having a lot of acquaintances; his friend added that women in particular “really want to get to know him.” He reported holding several jobs as an adult, including waiter, cook, and construction worker. He apparently held a number of briefer jobs as well, and described himself as “king of the misers—I don’t need a lot to get by.”

Mr. Stallworth reported a significant medical history, including “enlarged heart ventricles” that result in “pains in my heart and I pass out” at times. He was also diagnosed with asthma and prescribed a steroidal inhaler that “seemed to work well,” he said. He reported a history of two concussions, in 1993 (when he was thirty-three) and 1995, the latter incurred during a “fight with a bunch of people over alcohol.”

His psychiatric history included a report of hospitalization during college and diagnosis of manic depression, with a prescription for lithium (a mood-stabilizing psychotropic medication) that he said he “never took” because he “did not like the side effects.” He described one subsequent episode of manic behavior about two years later and two episodes of significant depression (in 1988 and 2000). He said that he did not take medication to treat any of these three episodes. He also described a history of symptoms consistent with ADHD, although he reported that he had never been diagnosed with or treated for this disorder.

The results of a standard psychological test of mental and emotional functioning, the MMPI-2, yielded a valid profile that was consistent with an open response style neither exaggerating nor minimizing unusual experience. Individuals with such profiles are often described as interacting with others easily, although prone to taking advantage of others. Such profiles are also consistent with substance-use problems—consistent with Mr. Stallworth’s description of using marijuana, alcohol, and cocaine excessively and using other drugs (e.g., quaaludes, LSD) when in college.

Testimony in the Federal Sentencing Hearing

Dr. A described the following recommendations for treatment: (1) a medical expert’s evaluation for appropriateness of psychotropic medication to treat

symptoms of bipolar disorder and ADHD; (2) job-related services that could assist Mr. Stallworth in finding and keeping a job, reducing the pressure on him to obtain money by antisocial means; and (3) skills-based training and counseling to address deficits in problem solving and decision-making. His amenability to such interventions was described as mixed. He has the intellectual capacity to benefit from interventions relying on verbal memory and abstract concepts, and he has acknowledged that psychotropic medication has helped reduce the intensity of his manic symptoms. But he also has a history of declining or discontinuing prescribed medication, so he would need to be persuaded that it was advantageous to him under the circumstances.

Dr. B noted Mr. Stallworth's long history of antisocial conduct as an adult and concluded that his report of experiencing bipolar symptoms was exaggerated and self-serving. He concluded that Mr. Stallworth does not experience a severe mental illness but has a personality disorder (antisocial personality disorder) co-occurring with substance misuse. Dr. B also observed that he was glib, impulsive, and versatile in his antisocial conduct. He was pessimistic about the prospect of rehabilitation, concluding that no conventional psychiatric interventions (including psychotropic medication) were likely to reduce his risk of future offending.

Questions for Consideration: Dr. A

1. Given that Dr. A was based in a local university and devoted much of his work to research and teaching, including conducting forensic evaluations but doing relatively little work treating patients, should he have been considered qualified to offer opinions about Mr. Stallworth's treatment needs and amenability?³¹¹
2. Were the psychological tests administered appropriate in this context?³¹²
3. Are the following relevant to (a) whether Dr. A should be qualified as an expert, (b) the credibility of his opinions but not his status as an expert, or (c) neither?

311. The most important question involves this expert's knowledge, training, and experience in conducting similar forensic evaluations. Other sources of relevant information include whether the expert provides treatment to this population (although, as discussed in this reference guide, clinical treatment provision is a separate and distinct role from forensic assessment), whether he has conducted research and scholarship in related areas, whether he has trained peers and trainees, and other considerations that might inform his competence in conducting such an evaluation and reporting the results.

312. The psychological tests administered in this case were selected for relevance and reliability. This is not always the case with experts who administer such testing. Judges who have questions about this should ask the expert to justify an indicated measure in terms of relevance and reliability.

- a. psychologist (not a medical expert)³¹³
 - b. licensure to practice as a psychologist³¹⁴
 - c. board certification³¹⁵
 - d. experience treating patients³¹⁶
 - e. current work in forensic evaluations, treatment provision, and treatment of justice-involved individuals³¹⁷
 - f. retained by the defense³¹⁸
4. Dr. A described Mr. Stallworth in a way that included both strengths and limitations, and suggested that certain kinds of interventions might be appropriate for targeting his symptoms. Does that provide an accurate depiction in this case?³¹⁹

Questions for Consideration: Dr. B

1. Does Dr. B's primary employment as a private practitioner who conducts forensic psychiatric evaluations, along with consultation and treatment of

313. Psychologists are generally appropriate experts in such matters. Considering expertise via voir dire might reveal justification for clearly accepting or rejecting the proposed expert on other grounds.

314. This is important. Licensure in a jurisdiction indicates that the appropriate regulatory board has determined that the psychologist has the necessary training and experience to deliver psychological services to clients. Forensic assessment is a kind of psychological service that involves interviewing, testing, and expertise in mental health. Unlicensed psychologists have not demonstrated that they have the needed training and experience to provide this service. Some psychologists who specialize in nonclinical areas neither qualify for licensure nor are appropriate to provide forensic assessments.

315. This is a demonstration that the psychologist has obtained a credential that requires competence in the area of specialization. Some psychological experts have not sought board certification but are nonetheless competent to provide forensic services.

316. This is important to the extent that it furthers expertise with such individuals and promotes a better understanding of how treatment works, and its limits. It is less important that the expert have an active clinical treatment practice; experts engage in many activities that further their competence. Treatment provision is one. Research, training, and consultation are others.

317. Experience providing services with justice-involved clients is important. Such services include areas described in the previous footnote.

318. Experts can agree or decline to work with an attorney, but they generally do not control whether they receive such offers. The proportion of cases in which one has been retained by one side is less important than demonstrating some history of conducting evaluations at the request of both defense and prosecution or defense and plaintiff. Also important is the proportion of cases in which the expert can describe providing an opinion that was not favorable to the retaining party. Such unfavorable opinions typically are not used in litigation, so they do not become part of the discoverable record. But an expert should be able to answer a question about this.

319. Whether the described strengths and limitations are accurate depends upon the evidence presented, of course. But strengths and limitations always exist, and reports and opinions that do not describe both should be viewed with some skepticism with respect to the evaluator's impartiality.

- patients in different psychiatric facilities, give him an “expertise advantage” in this case?³²⁰
2. Is the evaluation limited by the absence of standardized psychological testing?³²¹
 3. Are the following relevant to (a) whether Dr. B should be qualified as an expert, (b) the credibility of his opinions but not his status as an expert, or (c) neither?
 - a. psychiatrist (a medical expert)³²²
 - b. licensure to practice as a physician³²³
 - c. board certification³²⁴
 - d. experience treating patients³²⁵
 - e. current work in forensic evaluations, treatment provision, and treatment provision with justice-involved individuals³²⁶
 - f. retained by the prosecution³²⁷
 4. Dr. B described Mr. Stallworth without particular attention to strengths and by observing that he does not experience a severe mental illness, instead concluding that he has a personality disorder that is very difficult to treat, with co-occurring substance misuse. Does that provide an accurate depiction in this case?³²⁸

320. It provides evidence that much of his time is spent with patients. This has advantages and disadvantages. To the extent that an expert spends the great majority of his time with patients, and is not involved in other professional activities (e.g., research, scholarship, training), that expert may be very familiar with patient behavior but less familiar with the supporting science and practice literatures.

321. Psychiatrists rarely administer psychological testing, as training in this area is not typically provided during residency or fellowship. This means they may not be well informed about what testing could provide. But such experts usually take alternative approaches to appraising important domains in forensic assessment.

322. Psychiatrists with appropriate expertise in areas relevant to the forensic questions should be qualified as experts.

323. This is simply an indication that this individual is medically trained, which may have certain advantages considering the nature and facts of the case.

324. Board certification is expected more often as part of medical and specialty training. It is a reflection of such specialization.

325. Having a clinical practice adds to the credibility and experience of an expert. There should be a recognition, however, that the forensic-assessment role is distinct from the clinical-treatment role. For reasons discussed in this reference guide, experts who have treated the litigant should not then be providing a forensic assessment of that litigant.

326. Specialized experience with justice-involved individuals is important.

327. Unless there is an identifiable pattern of only conducting evaluations and testifying at the request of one side, the particular side that retained the expert in this matter should make little difference.

328. Whether this is accurate will depend on the pattern of the evidence provided by both experts. But describing a litigant without reference to strengths, or in one-sided terms, should create skepticism with respect to the expert's impartiality.

Glossary of Terms

actuarial assessment. A form of structured assessment in which scores on assessment instruments are associated with outcomes (linked most often through empirical research) without requiring a judgment from the evaluator. This is one of the two most accurate approaches to prediction—the other being structured professional judgment.

behavior therapy. A form of treatment for individuals with psychological symptoms that focuses on the influence of different stimuli on these symptoms and the individual's response expressed in their behavior.

blind spot bias. A form of cognitive bias in which there is limited awareness that the influences that contribute to cognitive bias in others also apply to oneself.

board certification. Certification of competence in a particular specialization by a specialty board in psychiatry or psychology. It differs from licensure in its requirement of more specific measures of competence.

clinical versus forensic contexts. The delineation of the ways in which the roles of clinical treatment provision versus forensic mental health assessment differ. They include purpose, client, relationship, voluntariness, confidentiality, consent, response style, sources of information, pace, settings, report, and testimony.

cognitive behavioral therapy. A form of treatment of individuals with psychological symptoms that focuses on the influences of different stimuli, and individuals' thoughts and beliefs, on these symptoms, and on the individuals' responses expressed in their behavior, thoughts, and beliefs.

cognitive biases. Systematic influences on decisions made under conditions of uncertainty that may result in inaccuracy, illogic, and inconsistency with objective evidence. Such influences often operate outside the awareness of the decision-maker. For examples, see confirmation bias, retention bias, and blind spot bias (all in this glossary).

confirmation bias. The tendency to identify, interpret, and remember information in a fashion that is consistent with (or confirms) existing values or decisions.

construct validity. A form of validity that refers to a measure's accuracy in assessing what it is intended to assess.

contemporaneous evaluations. Expert evaluations appraising a party's relevant thinking, behavior, and capacities at present.

content validity. A form of validity referring to the representativeness of a measure's sample (the items used to provide empirical evidence about the measure) of the outcome (e.g., behavior, skill) being measured.

current-state evaluations. See contemporaneous evaluations.

- decisional capacity.** Ability to make a legally relevant decision in an acceptable way, which may be influenced by cognitive capacity and symptoms. Capacity references the process of decision-making rather than the final decision itself.
- diagnosis.** Categorization of mental disorder(s), most often using the current version of the *Diagnostic and Statistical Manual of Mental Disorders* published by the American Psychiatric Association.
- dialectical behavior therapy.** A form of treatment of individuals with psychological symptoms that focuses on helping people regulate their thoughts and feelings through identifying triggers and using skills to help them cope with problematic events and their responses to the events.
- DSM-5-TR.** *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, Text Revision.* The most recent diagnostic manual published by the American Psychiatric Association.
- forensic mental health assessment.** An evaluation conducted by a mental health expert, typically a psychiatrist or a psychologist, with the purpose of informing the court about a party's symptoms, characteristics, and capacities relevant to a legal question involving that party, and/or assisting the retaining attorney in providing accurate evidence regarding that party.
- forensic psychiatric assessment.** Forensic mental health assessment conducted by a psychiatrist.
- forensic psychological assessment.** Forensic mental health assessment conducted by a psychologist.
- functional legal capacities.** A party's abilities to think and behave on tasks associated with the legal question. Along with psychological symptoms and characteristics, functional legal capacities are appraised and considered by the expert in evaluating a party on a specific legal question.
- intellectual disability.** Formerly called mental retardation, referring to a disability characterized by significant limitations in both intellectual functioning and adaptive behavior, with onset before age twenty-two.
- licensure.** Recognition by a regulatory board in a given jurisdiction that an individual has the requisite training and experience in a particular discipline to be authorized to function independently in the designated professional role.
- performative capacity.** Ability to carry out a legally relevant task, which may be influenced by cognitive capacity and symptoms.
- predictive validity.** A form of validity that refers to the capacity of a test or measure to accurately predict a specified future outcome.
- prospective evaluations.** Expert evaluations appraising a party's likely future mental states, cognitive function, behavior, and capacities.

psychiatrist. Mental health specialist trained in medical school and psychiatry residency. Some subspecialists also train in fellowships following residency.

psychologist. Mental health specialist who has obtained a doctoral degree and completed a clinical internship. Some specialists also train in fellowship following internship.

reliability. The scientific definition of reliability refers to the consistency of a given measure—the extent to which the same results are obtained and are not affected by error across different measurements or assessors. The legal definition of reliability combines both scientific reliability (measurement consistency) and scientific validity (measurement accuracy) into a single concept that refers to overall accuracy.

response style. The tendency of a person to provide self-referential information that is accurate (within limits of memory) versus deliberately distorted. Such distortions can include overreporting/exaggeration/fabrication of symptoms (called malingering or exaggeration), underreporting or denial of symptoms (called minimization or denial), or refusing to provide relevant information (through responding irrelevantly on psychological testing, failing to provide much information when questioned, or refusing to talk with an evaluator).

retention bias. A form of cognitive bias in forensic psychologists and forensic psychiatrists that involves identifying, interpreting, and remembering information in a way that favors the party retaining the expert.

retrospective evaluations. Expert evaluations appraising a person's relevant thinking, behavior, and capacities at a specified previous time. In some cases, it cannot involve direct contact with that person (e.g., testamentary or contractual capacity of a deceased individual).

risk-need-responsivity. A leading theory of correctional classification that includes risk of reoffending, needs for intervention for risk-relevant deficits, and likelihood of responding favorably to interventions, both generally (using empirically supported procedures) and specifically (with interventions consistent with the individual's capacities and learning style).

social-skills therapy. A form of behavior therapy used primarily with individuals with severe mental illness or developmental disability, focusing on learning and improving important social skills that help such individuals interact more effectively with others.

structured assessment techniques. Standardized interviews or data-gathering protocols designed to obtain clear responses to specific questions.

structured professional judgment. A form of semistructured assessment in which evaluators gather information in predefined relevant areas and, in light of that information, apply professional judgment regarding the likelihood

of a predefined outcome. This is one of the two most accurate approaches to prediction—the other being actuarial assessment.

validity. Being logically or factually accurate, an important consideration in science that refers to whether a measurement represents what it is supposed to measure. There are different kinds of validity. See construct validity, content validity, and predictive validity.

References on Mental Health Diagnosis and Treatment

- American Psychiatric Association, *The American Psychiatric Association Practice Guidelines for the Psychiatric Evaluation of Adults* (2016).
- American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders (DSM-5-TR)* (5th ed. text rev. 2022).
- American Psychiatric Publishing *Textbook of Psychiatry* (Laura W. Roberts ed., 7th ed. 2019).
- APA Handbook of Forensic Neuropsychology (Shane S. Bush et al. eds., 2017).
- Kaplan and Sadock's *Comprehensive Textbook of Psychiatry* (Benjamin J. Sadock et al. eds., 10th ed. 2017).
- Alan F. Schatzberg & Charles DeBattista, *Schatzberg's Manual of Clinical Psychopharmacology* (9th ed. 2019).
- Stephen M. Stahl, *Essential Psychopharmacology: The Prescriber's Guide* (7th ed. 2020).

References on Mental Health and Law

- APA Handbook of Forensic Neuropsychology (Shane S. Bush et al. eds., 2017).
- Paul S. Appelbaum, *A Theory of Ethics for Forensic Psychiatry*, 25 J. Am. Acad. Psychiatry & L. 233 (1997).
- Clinical Assessment of Malingering and Deception (Richard Rogers & Scott Bender eds., 4th ed. 2018).
- Deborah Giorgi-Guarnieri et al., *American Academy of Psychiatry and the Law Practice Guideline for Forensic Psychiatric Evaluation of Defendants Raising the Insanity Defense*, 30 J. Am. Acad. Psychiatry & L. S1 (2002).
- Thomas Grisso, *Evaluating Competencies: Forensic Assessments and Instruments* (2d ed. 2002).
- Gisli H. Gudjonsson, *The Psychology of Interrogation and Confessions* (2003).
- Thomas G. Gutheil & Paul S. Appelbaum, *Clinical Handbook of Psychiatry and the Law* (5th ed. 2020).
- Gary B. Melton et al., *Psychological Evaluations for the Courts: A Handbook for Mental Health Professionals and Lawyers* (4th ed. 2018).
- Mental Disorder, Work Disability, and the Law (Richard J. Bonnie & John Monahan eds., 1997).
- John Monahan, *The Scientific Status of Research on Clinical and Actuarial Predictions of Violence*, in *Modern Scientific Evidence: The Law and Science of Expert Testimony* (David L. Faigman et al. eds., 2007).
- Robert D. Morgan et al., *Treating Offenders with Mental Illness: A Research Synthesis*, 30 L. Hum. Behav. 1, 37–50 (2012).

Reference Manual on Scientific Evidence

- Douglas Mossman et al., *American Academy of Psychiatry and the Law Practice Guideline for the Forensic Psychiatric Evaluation of Competence to Stand Trial*, 35 J. Am. Acad. Psychiatry & L. S3 (2007).
- Michael L. Perlin, *Mental Disability Law: Civil and Criminal* (3d ed. 2016).
- Retrospective Assessment of Mental States in Litigation: Predicting the Past (Robert I. Simon & Daniel W. Shuman eds., 2002).
- Christopher Slobogin, *Proving the Unprovable: The Role of Law, Science, and Speculation in Adjudicating Culpability and Dangerousness* (2006).

Reference Guide on Engineering

CHAOUKI T. ABDALLAH, BERT BLACK, AND EDL SCHAMILOGLU

Chaouki T. Abdallah, Ph.D., is Professor, School Electrical & Computer Engineering, and Former Executive Vice President for Research, Georgia Institute of Technology.

Bert Black, M.S., J.D., is a partner with Schaefer Halleen, LLC.

Edl Schamiloglu, Ph.D., is Distinguished Professor, Department of Electrical & Computer Engineering, University of New Mexico.

CONTENTS

Introduction, 1355

What Is Engineering and How Does It Differ from Science?, 1357

 Engineering and Science as Complementary Fields, 1359

 The Scientific Method, 1359

 The Engineering Design Process, 1360

Who Is an Engineer?, 1363

 Academic Education and Training, 1363

 Licensing, Registration, Certification, and Accreditation, 1365

How Do Engineers Think, Make, Test, and Evaluate?, 1368

 Structured Systems-Level Models and Thinking, 1368

 Adeptness at Designing Under Constraints, 1369

 Understanding Tradeoffs, 1370

Engineering Systems, 1371

 Systems and Systems of Systems, 1371

 Complex Systems, 1372

 When Systems Fail, 1373

 Computers, Artificial Intelligence, and Machine Learning, 1375

Analysis of the Case Law Based on an Understanding of Engineering and How It Differs from Science, 1376

 Products Liability Cases, 1377

 Design Defects, 1378

 Reasonable alternative design, 1379

 Is testing required to prove existence of a defect or the validity of an alternative design?, 1380

 Manufacturing Defects, 1385

 Warning Defects, 1388

 Causation, 1390

 Duty of Care, 1392

 Non-Products Liability Cases, 1393

Reference Manual on Scientific Evidence

Contract and Construction Litigation Cases, 1393
Intellectual Property Cases, 1395
Engineers in copyright cases, 1395
Engineers in trademark cases, 1396
Engineers in patent cases, 1396
Claim construction, 1398
Infringement, 1400
Validity, 1401
Engineers in trade secret cases, 1403
Environmental/Land Use Cases, 1404
Failure Analysis Cases, 1405
Testing Cases, 1405
Conclusion, 1406
Glossary of Terms, 1407

TABLES

1. List of Engineering Fields, 1361
2. Steps of Scientific Method Versus Engineering Design Process, 1362

Introduction

Courts have a long history resolving disputes about the admissibility of opinion testimony from engineers proffered as expert witnesses.¹ In some cases, however, background guidance on how engineers think and go about doing their work, and on who qualifies as an engineer, would be helpful to both judges and lawyers. This reference guide aims to fill that need.² It builds upon similar guides, in the second³ and third⁴ editions of the *Reference Manual on Scientific Evidence*, that were added in the wake of the Supreme Court’s 1993 decision in *Daubert v. Merrell Dow Pharmaceuticals, Inc.*⁵ and its 1999 decision in *Kumho Tire Co., Ltd. v. Carmichael*.⁶

In *Daubert*, as discussed below, the Court relied heavily on the philosophy of science to provide guidance for gatekeeper judges who must resolve disputes about what should or should not count as valid and admissible scientific evidence,⁷ but also explained how the actual practice of science is more nuanced and complex than the textbook scientific method. “Scientific conclusions are subject to perpetual revision . . . The scientific project is advanced by broad and wide-ranging

1. *Folkes v. Chadd*, 3 Doug. 157, 99 Eng. Rep. 589 (1782), an eighteenth-century English case often cited as the first in which a court allowed expert opinion testimony, involved engineering experts who provided conflicting explanations for what had caused a harbor to become clogged with silt. See Tal Golan, *Revisiting the History of Scientific Expert Testimony*, 73 Brook. L. Rev. 879, 886–904 (2008).

2. Based on a Westlaw search of engineering expert cases decided during the 18 months from January 2021 through June 2022 (discussed in detail below, in the section titled “Analysis of the Case Law Based on an Understanding of Engineering and How It Differs from Science”), federal courts currently decide about 240 cases a year involving engineering experts, of which only 134 involve the kind of complex engineering testimony on which this reference guide focuses. In just over 20% of the 240 cases, the experts were accident investigators, an area with which courts have become quite familiar and comfortable after a history going back more than a hundred years. See, e.g., *Grand Trunk Ry. Co. of Canada v. Ives*, 144 U.S. 408, 412 (1892) (expert testified that the distance traveled by a train after colliding with a buggy indicated “it must have been going at the rate of 25 or 30 miles an hour”); *Shugart v. Atlanta, K. & N. Ry.*, 133 F. 505, 507–08 (6th Cir. 1904) (allowing conflicting expert testimony about whether condition of railroad tracks had caused a derailment). Fifteen percent of the cases involved insurance coverage issues, and about 7% involved matters like employment claims or criminal law issues that don’t really implicate complicated engineering questions. Again, this reference guide focuses on the more complex engineering issues that arise in the remaining 58% of the cases, the majority of which involved products liability claims.

3. Henry Petroski, *Reference Guide on Engineering Practice and Methods*, in *Reference Manual on Scientific Evidence* 579–624 (2d ed. 2000).

4. Channing R. Robertson, John E. Moalli, & David L. Black, *Reference Guide on Engineering*, in *Reference Manual on Scientific Evidence* 897–959 (3d ed. 2011).

5. 509 U.S. 579 (1993).

6. 526 U.S. 137 (1999).

7. *Daubert*, 509 U.S. at 593–95; Liesa L. Richter & Daniel J. Capra, *The Admissibility of Expert Testimony*, in this manual at 5–13 (hereinafter Richter & Capra).

consideration of a multitude of hypotheses, for those that are incorrect will eventually be shown to be so, and that in itself is an advance.”⁸

Daubert also provided a “five-factor, nondispositive, nonexclusive, ‘flexible’ test” for trial courts to use when determining the validity of scientific evidence, as required by Rule 702.⁹ The five factors, generally referred to as the “*Daubert* factors,” are:

1. whether the technique or theory can be or has been tested;
2. whether the theory or technique has been subject to peer review and publication;
3. the known or potential rate of error;
4. the existence and maintenance of standards and controls; and
5. the degree to which the theory or technique has been generally accepted in the scientific community.

In the immediate aftermath of *Daubert*, there was uncertainty about whether its gatekeeping mandate applied to all expert testimony or just to scientific experts, and if it did apply more broadly, whether the five factors should be used for non-scientific experts.¹⁰ *Kumho Tire* resolved the uncertainty about broader applicability of *Daubert* by extending judicial gatekeeping to all expert evidence, but provided no analogous general guidance or list of factors to consider when evaluating the engineering expert testimony (from a tire failure expert) at issue in that case. Instead, the Court’s decision focused on the specific evidence the trial court had excluded and concluded the trial court had not abused its discretion. The Court’s reluctance to delve into general guidance or factors for fields of expertise other than science reflects the fact that except for science (the field of human endeavor focused on expanding our knowledge of the natural world), there exists no extensive literature on the philosophical underpinnings for the knowledge experts may develop or acquire. For engineering, however, there is a wealth of practical guidance that forms the basis for this reference guide, which addresses three fundamental questions: (1) What is engineering, and how does it differ from science? (2) Who is an engineer? and (3) How do engineers think? In particular, how do they create, test, and evaluate designs? As explained in the

8. *Daubert*, 509 U.S. at 597. See also Michael Weisberg & Anastasia Thanukos, *How Science Works*, in this manual, which makes clear that textbook and media versions of the scientific method (hypothesis, followed by experimental testing, followed by publication, followed by acceptance and consensus in the scientific community) is far too linear and simple, and does not reflect how science actually works in practice. *How Science Works*, at 50–56.

9. For more on *Daubert*, Rule 702, and admissibility, see Richter & Capra, *supra* note 7.

10. Compare *Watkins v. Tel Smith, Inc.*, 121 F.3d 984, 991 (5th Cir. 1997) (“the nonexclusive list of factors relevant under *Daubert* to assessing scientific methodology . . . are also relevant to assessing other types of expert evidence”) with *McKendall v. Crown Control Corp.*, 122 F.3d 803, 806 (*Daubert* factors “are relevant only to testimony bearing on ‘scientific’ knowledge”).

following section, the creation and analysis of designs is the central feature of engineering practice, and it is the primary focus of this reference guide.

What Is Engineering and How Does It Differ from Science?

Engineers do often use or apply scientific knowledge, but that does not mean engineering is in any way subsidiary to science. Indeed, the practice of engineering pre-dates science by thousands of years. There is, however, an interplay between the fields, which means judges sometimes may have to decide whether to apply some or all of the *Daubert* scientific evidence criteria or different engineering criteria when determining the admissibility of an engineer's testimony.¹¹ Thus, the necessary starting point for this reference guide is a discussion of how the two fields differ and how they relate to each other. What, then, distinguishes engineering from other fields, especially science? Theodore von Kármán, an eminent aerospace engineer (also a mathematician and physicist) once explained that “scientists study the world as it is, engineers create the world that never has been.”¹²

More precisely, engineering and science differ in three important ways:

1. Historically, many engineering products and technologies have preceded scientific understanding,¹³ and while this remains true today, the advent of modern science often allows engineers to use scientific discoveries as part of their tool kit when they create new designs.¹⁴ In no event, however,

11. For example, if called upon as an expert witness to determine the cause of an accident or the failure of a product, an engineer might take various approaches. This reference guide is most applicable when an engineering expert bases such a determination on the analysis of a design. *See, e.g.,* *Baugh v. Cuprum S.A. de C.V.*, 845 F.3d 838 (7th Cir. 2017) (in which a mechanical engineer relied on calculations based on centuries-old mathematical principles to determine the cause of a ladder failure). If, on the other hand, an engineering expert formulates a hypothesis that requires testing or other empirical corroboration, some of the *Daubert* factors used for evaluating science might be more applicable. *See, e.g.,* *Nease v. Ford Motor Co.*, 848 F.3d 219 (4th Cir. 2017) (in which the plaintiff's expert hypothesized that debris in a cable housing had caused an accelerator to stick but offered no analytical explanation and no testing of his hypothesis).

12. Petroski, *supra* note 3, at 579. Theodore von Kármán was the National Medal of Science recipient in 1962.

13. William S. Hammack & John L. Anderson, *Working in the Penumbra of Understanding*, *Issues in Sci. & Tech.*, Feb. 16, 2022, available at <https://perma.cc/48H5-WUCN>.

14. One good example is the use of discoveries by the Nobel Laureate Jacques Monod about how bacteria reproduce as part of the design process for wastewater treatment plants. Alexandru Braha & Ferdinand Hafner, *Use of Monod Kinetics on Multi-Stage Bioreactors*, 19 *Water Rsch.* 1217 (1985), [https://doi.org/10.1016/0043-1354\(85\)90174-5](https://doi.org/10.1016/0043-1354(85)90174-5).

should engineering be considered a lesser field or fully dependent on science. Engineering is not just “applied science.”

2. Engineering begins with a goal and proceeds to the design and creation or manufacture of an object or a process to achieve the goal, while science begins with a question and strives to find an answer.
3. Engineers are comfortable proceeding toward their design goal in the presence of uncertainty, which they allow for in various ways, including safety factors that account for approximations in making design calculations, and variability and uncertainty in the execution of designs. Engineering has also been described as a “set of tradeoffs,” and tradeoffs are the engineers’ way of adapting to constraints. In contrast, for scientists, reduction of uncertainty by increasing our knowledge and understanding of the natural world is the central objective.

Science and engineering do sometimes get lumped together, or even mistaken for each other, and this conflation has prompted many efforts to differentiate and contrast them.¹⁵ Petroski in the second edition and Robertson et al. in the third edition of the reference manual presented detailed contrasts between the two fields. Perhaps the most succinct description was given by Gilbert Ryle: science is about “knowing that” while engineering is about “knowing how.”¹⁶ Our intent here is to describe and contrast two major concepts: the scientific method and the engineering design process, and then to explain how the design process works in practice. The revisions to Rule 702 make this inquiry even more important than it would be otherwise. As explained in *The Admissibility of Expert Testimony*, in this manual, one goal of the changes is to make sure experts “stay within the bounds of what can be concluded from a reliable application of [their] basis and methodology.”¹⁷ While prompted primarily by concerns about forensic science experts,¹⁸ the revisions relate to all experts, and if the reliable use of the process for creating and analyzing designs is at issue, application of scientific criteria and the *Daubert* factors often will not be the best way to determine admissibility. The question should not be reduced to whether a witness is called an engineer or not, but whether the witness is formulating and testing hypotheses or creating or evaluating designs (based on established methods and the witness’s experience and training).¹⁹

15. Billings Clinic, *The Scientific Method vs. Engineering Design Process*, available at <https://perma.cc/H6CZ-98TE>.

16. Gilbert Ryle, *Knowing How and Knowing That*, Proceedings of the Aristotelian Society, XLVI (in Collected Essays 1929–1968, Collected Papers Vol. 2, pp. 212–25).

17. Richter & Capra, *supra* note 7, at 23, in this manual, quoting Advisory Committee’s note to 2023 amendment to Fed. R. Evid. 702.

18. *Id.*

19. The Supreme Court’s observation in *Kumho Tire* about engineering resting on a scientific foundation may have inadvertently created some confusion on this point. In one sense it could be

We shall see below in the section titled “How Do Engineers Think, Make, Test, and Evaluate” that there are three essential properties of the engineering mindset: the ability to see a structure where there’s nothing apparent; adeptness at designing under constraints; and understanding tradeoffs.

Engineering and Science as Complementary Fields

Engineering is actually a much older field of human endeavor than science. The creation of tools during the Stone Age involved the conception and execution of designs and took place long before humans began any scientific effort to understand nature. Thus, many engineering designs necessarily predate science, even in more recent times. A classic example is the development of the steam engine, which triggered the science of thermodynamics.²⁰ Contrary to the popular notion that science (first) discovers, and engineering (then) applies, the “relationship between science and engineering is . . . complementary, synergistic, and essential. Scientific practice and knowledge offer engineers . . . heuristics that work better than those based . . . on trial and error; but this scientific knowledge does not explain how to design or create an artifact or a system.”²¹ Moreover, scientists “use the products of engineering to investigate and discover.”²²

The Scientific Method

Scientists strive for a common explanation for a phenomenon they are investigating. They may start with competing hypotheses, but the scientific method should eliminate most, if not all, and yield a consensus on which explanation is most likely correct. For example, although a few scientists still dispute the validity of evolution as the explanation for how a huge variety of species developed, there is a broad consensus that it provides the most likely explanation today, keeping in mind that future experimentation may invalidate the current consensus.

Although science does not proceed in a linear process from hypothesis to testing to publication and acceptance, it is quintessentially about

taken to mean the criteria for scientific validity and reliability should be used to evaluate all engineering testimony, but the far better interpretation is that engineers sometimes use scientific discoveries in the course of creating designs. If the scientific underpinning were not well accepted and reliable, then the engineering use of it could be called into question, but that is not fundamentally related to the engineering design process.

20. Petroski, *supra* note 3, at 580; Robertson et al., *supra* note 4, at 902. *See also* Hammack & Anderson, *supra* note 13.

21. *See* Hammack & Anderson, *supra* note 13.

22. *Id.*

figuring out what predictions are generated by a hypothesis and comparing those to evidence. Hypotheses are supported when the evidence matches predicted observations and are contradicted when they do not match. Observations are often made with the help of sophisticated tools and may be analyzed with statistical techniques to discern patterns. Other observations are entirely qualitative and nonnumeric.²³

This is consistent with Occam's razor (or the principle of parsimony), which states that the explanation that requires the fewest assumptions is usually correct.

Thus, one distinguishing characteristic of scientific explanations is that they have to be testable. The traditional conception of science formulated by philosophers like Karl Popper would have falsifiability as the foundational principle of science, but the current view is that science cannot prove any idea to be false (or true for that matter). Furthermore, "reviewing the logical arguments presented in any scientific journal will reveal that evidence can and does play a role in supporting particular hypotheses over others, not just in ruling some ideas out, as implied by the doctrine of falsification."²⁴ Testing and modification of even well-established scientific explanations is a more realistic basis for determining the validity of science.

The classic example is how Newtonian dynamics withstood the test of time, until experimental evidence showed its limitations, paving the way for Einstein's theory of special relativity. Matters did not end there, however. Experiments revealed limitations in Einstein's theory, which led to quantum mechanics—and another explanation, more consistent with observations and experimental data, may yet emerge. In fact, a frequently misunderstood and miscommunicated concept is that scientific truths are absolute and invariant—in fact, they're accepted as correct until they become inconsistent with what's revealed by further experimentation. Moreover, and as described by Pitt, scientific knowledge is produced by researchers "exploring the domain of a *theory* who aim to provide an account of the relations among the objects and processes of that domain, an account which provides the basis for an explanation of phenomena generally observed or detected in another domain." Thus, "scientific knowledge is theory-bound" and as theories are modified, scientific knowledge (i.e., consensus) also changes.²⁵

The Engineering Design Process

Engineering is a vast field of practice, consisting of many recognized subfields and specializations, such as electrical engineering, mechanical engineering, civil engineering, systems engineering, aeronautical engineering, and many more. The

23. See Michael Weisberg & Anastasia Thanukos, *How Science Works*, p. 82, in this manual.

24. *Id.*

25. Joseph C. Pitt, *What Engineers Know*, *Techné* 5 (3): 19 Spring 2001, available at <https://perma.cc/GA4X-5Q3E>.

Table 1. List of Engineering Fields

Aeronautical/aerospace/astronautical engineering	Industrial engineering
Agricultural engineering	Marine engineering
Bioengineering/biomedical engineering	Materials and metallurgical engineering
Chemical engineering	Mechanical engineering
Civil, including architectural/sanitary engineering	Mining and geological engineering
Computer engineering	Nuclear engineering
Electrical and electronics engineering	Petroleum engineering
Environmental engineering	Sales engineering

National Academy of Engineering Report, *Understanding the Educational and Career Pathways of Engineers (2018)*,²⁶ lists 16 engineering fields,²⁷ shown in Table 1.

It should be noted that the Accreditation Board for Engineering and Technology, Inc. (ABET) maintains a list of engineering programs that they review for accreditation,²⁸ and this list also includes fire protection engineering, geological engineering, optical engineering, software engineering, and systems engineering.

To further illustrate the breadth of engineering, the IEEE (Institute of Electrical and Electronics Engineers), one of the world’s largest professional organizations, comprises 39 technical societies ranging from Aerospace and Electronic Systems to Vehicular Technology.²⁹ Engineering specializations are ever changing as societal needs evolve along with new discoveries, inventions, and technologies. Specialties such as nuclear engineering and biomedical engineering are relative newcomers, while civil engineering was practiced by early civilizations.

All engineering disciplines, however, share certain common practices and processes for creating designs and solving problems. The engineering design process is the touchstone for all engineers, much as the scientific method is the touchstone for all scientists. Thus, understanding how engineering design is distinct from the scientific method is important to the Rule 702 gatekeeping task. While engineers may differ on their approach to solving a problem, “given certain assumptions about the contingencies involved—it is not the case that two

26. National Academy of Engineering, *Understanding the Educational and Career Pathways of Engineers (2018)*, <https://doi.org/10.17226/25284>.

27. The list includes sales engineering, defined in the standard occupational classification system used by the National Science Foundation (NSF) as those who “sell business goods or services, the selling of which requires a technical background equivalent to a baccalaureate degree in engineering.”

28. ABET, *Criteria for Accrediting Engineering Programs, 2023–2024*, available at <https://perma.cc/JZ2G-QNMU>.

29. IEEE Societies, available at <https://perma.cc/6N8T-J3JX>.

engineers similarly educated and experienced could be armed with sufficiently different perspectives that they would flat out contradict each other.”³⁰ This makes engineering testimony less about discovering or converging on “the one truth” and more about adherence to process and practice. Although neither the traditional conception of the scientific method nor any effort to precisely define the engineering process fully captures all the nuances of what happens in either field, a comparison of traditional simplified views of both science and engineering does help to highlight how the two fields differ. Table 2 shows this comparison.

Table 2. Steps of Scientific Method Versus Engineering Design Process

	Scientific Method	Engineering Design Process
Step 1	State your question	Define the goal
Step 2	Conduct background research	Conduct background research
Step 3	Construct a hypothesis	Specify requirements and constraints
Step 4	Design experiment	Consider possible design options; evaluate and choose among them
Step 5	Test hypothesis via experiment	Develop a solution
Step 6	Analyze results	Verify performance, safety, and economics through analysis, and/or testing, and/or comparison with existing solutions
Step 7	If results align fully with hypothesis, communicate results through publication	If solution fully meets requirements, communicate it to those who will execute the design
Step 8	If not, go back to step 3	If not, go back to step 4, 5, or 6 as appropriate

In addition to the points listed in Table 2, another salient difference between science and engineering is that engineered products are intended to interact with human users or as part of an engineered system. It is not enough that a phone or a bridge function properly in the controlled conditions of a laboratory or in simulation. Phones must work despite the day-to-day abuse by their users, and bridges have to withstand various weather and congestion conditions and function properly under those conditions. Engineered products must also be safe, ergonomically designed, and aesthetically pleasing, as well as functional. It is indeed the various requirements and constraints that make engineering a creative synthesis of art (idea creation), scientific knowledge of how the world works, and accumulated design experience. Science, on the other hand, is more about explanatory theories than about practical application of those theories.

30. Joseph C. Pitt, *What Engineers Know*, Techné 5 (3): 17 Spring 2001, available at <https://perma.cc/GA4X-5Q3E>.

It has been said that “to engineer is to contrive,” and as we described earlier, practical needs have often led to engineering solutions that preceded mathematical and scientific understanding. Engineering is also omnipresent but largely invisible. From everyday products such as household appliances and cars, to large systems such as power plants, engineering products and systems are often noticed only when they fail to perform as expected, leading to possible loss of property, money, or even lives. In addition, at the heart of many engineering designs is the concept of scale; engineering products are in general (exceptions include military systems) meant to be used by many users of varying sophistication and expertise.

Who Is an Engineer?

This section describes who may qualify as an engineer for the purpose of providing expert testimony. An academic degree is only the starting point and may not even be a requirement in some cases. The type of work proposed experts do on a day-to-day basis is usually more important in judging their qualifications.

Academic Education and Training

An earned bachelor’s degree in an accredited engineering curriculum³¹ is generally sufficient to enter the professional workplace and begin to solve a wide variety of problems. It is less so the case for students who graduate with degrees in the basic sciences such as physics, chemistry, and biology, or in mathematics. Typically, but not always, basic science students will go on to earn graduate degrees. Of course, students who have earned an engineering degree also sometimes continue to the master’s or even doctorate level of study. In 2021, U.S. colleges and universities awarded approximately 146,233 bachelor’s degrees, 63,319 master’s degrees, and 12,403 doctoral degrees in all areas of engineering.³²

31. Accreditation is performed by ABET (Accreditation Board for Engineering and Technology). Founded in 1932 as the Engineer’s Council for Professional Development (ECPD), it was later renamed ABET. In the United States, accreditation is a nongovernmental, peer-review process that ensures the quality of the postsecondary education that students receive. Educational institutions or programs volunteer to undergo this review periodically to determine if certain criteria are being met. ABET accreditation is assurance that a college or university program meets the quality standards established by the profession for which it prepares its students. The quality standards that programs must meet to be ABET-accredited are set by the ABET professions themselves. This is made possible by the collaborative efforts of many different professional and technical societies. These societies and their members work together through ABET to develop the standards, and they provide the professionals who evaluate the programs to make sure that the programs meet those standards.

32. Engineering & Engineering Technology by the Numbers, ASEE (2021), <https://perma.cc/5YXZ-HR2Y>.

One can think of the educational process as providing engineering students with a tool kit from which they select “tools” that enable them, either individually or in teams, to participate in designing products or processes to address specific needs or problems. Because recently graduated engineering students are educated (i.e., have only classroom knowledge) as opposed to being trained (i.e., taught how to put classroom knowledge to practical use), one can never be quite sure how they will choose to use their tools in practice. Also, once in practice they may add to their tool kit through experience, training, or additional education.

The importance of experience and training after graduation means that people without a formal engineering degree may become quite effective and successful engineers. Indeed, some of the confusion between science and engineering stems from the fact that people with a scientific background sometimes work as engineers and vice versa. The distinction between scientists and engineers is quite recent. For example, over 500 years ago, Leonardo da Vinci was at once an accomplished engineer and scientist, as well as a renowned artist. A century later, Christopher Wren was an engineer, an architect, a mathematician, an astronomer, and a physicist. Today, people tend to work in more siloed professions, but crossover between science and engineering continues.³³

Without knowing how an engineer or scientist will use the tool kit acquired in college, and to what extent it will be replenished or modified as time goes on, it is not possible to predict what any individual might do to shape her or his career as time passes. There is a great deal of truth to the notion of learning on the job. Indeed, as one’s career unfolds, the number of opportunities expands, and with that come additional skills and an ever-increasing ability to make wise and informed choices and decisions. This ongoing learning is made even more critical with the increasing adoption of machine learning and artificial intelligence (AI) tools in engineering designs. There is no universal path to becoming an engineer, and being an engineer affords one the opportunity to continually remodel oneself as new and unexpected problems and challenges become evident.

33. In the twentieth century, Buckminster Fuller seamlessly combined elements of geometry (“science”), structures (“engineering”), and architecture (“art”) to conceive and develop an entirely new approach to architectural design. Architects Norman Foster and Frank Gehry seized on recent advances in computer science and engineering to develop innovative platforms for architectural design that paved the way for radical changes in structural and visual renderings. Striking examples include the Guggenheim Museum in Bilbao, Spain; the Walt Disney Concert Hall in Los Angeles; the Experience Music Project in Seattle; the London City Hall; the Beijing Airport; and the Reichstag in Berlin. Other examples include the famous scientist Linus Pauling (Nobel Prizes in Chemistry and Peace), who was trained as an engineer, and Albert Michelson (first American to win a Nobel Prize in science), whose work designing instruments to measure the speed of light was more engineering than science.

And so it is with the passage of time that the “title” of one’s degree becomes an increasingly murky description of who one is and what one does. This is why it is critical that titles do not overshadow the actual context of a degree when evaluating whether an “engineer” is testifying within her or his realm of expertise and the experience base at hand (i.e., an expert’s title may not reflect the knowledge the expert can provide). Thus, an assessment of an expert witness’s qualifications to testify on an engineering matter needs to be done on a case-by-case basis. The assessment must be based on the witness’s education, years of experience as a practicing engineer, and prior experience and qualification as an expert witness. There also are examples of experts qualified to testify despite little if any formal education in engineering and despite having no engineering background.³⁴

Licensing, Registration, Certification, and Accreditation

After graduation from college, some practicing engineers may go on to obtain professional engineer (PE) licensure. Licenses are required for engineering professionals in all 50 states and the District of Columbia if their services are *offered directly to the public and they would affect public health and safety*.³⁵ To become licensed, in most states, engineers must have a college degree from an ABET-accredited engineering college or university, work under a PE for at least four years, pass two intensive competency exams,³⁶ and submit a license application to their state licensing board. Each state has its own set of conditions for taking the two exams. Each state engineering board has its own pricing, educational requirements, and license procedures. Most boards require candidates to be enrolled in or have completed an ABET-accredited engineering degree. Some states let students from other approved or even non-accredited colleges take the

34. See, e.g., *Guay v. Sig Sauer, Inc.*, 610 F. Supp. 3d 423, 427–32 (D.N.H. 2022) (former fire-arms instructor and armorer found qualified to testify on design of pistol even though he was not an engineer and had done no work on the design or manufacturing of guns).

35. In some states, businesses generally cannot offer engineering services to the public, or cannot have a name that implies that they do so, unless they employ at least one PE. For example, New York requires that the owners of a company offering engineering services be PEs. See <https://perma.cc/2WWD-9E56>.

36. The Fundamentals of Engineering (FE) exam, formerly called the Engineer in Training test, is typically taken during an undergraduate student’s last year in college. People who pass this examination and graduate with an engineering degree are called engineering interns (EIs) or engineers in training (EITs). They become eligible to take the second test, called the Principles and Practice of Engineering examination, after they’ve accumulated enough experience, typically four years of doing engineering work.

exam. However, this isn't always the case. One should make sure to determine the relevant state requirements.³⁷ To retain their licenses in most states, PEs must continually maintain and improve their skills throughout their careers. Obtaining a PE license may be critical in some fields, such as civil engineering,³⁸ but is less so in others. PEs have the authority to certify documents such as reports, drawings, or calculations by signing and sealing (or "stamping") them, thus attesting to the accuracy and correctness of the documents and taking legal responsibility for them.

Many engineering professionals do not seek a PE license because their services are not offered directly to the public or they have no need to certify engineering documents. Whether an individual is licensed as a PE is neither sufficient nor necessary to establish her or his competency as an engineer. Furthermore, the two examinations required for licensure test only for knowledge gained and assimilated at the undergraduate level. It is therefore common for professors of engineering not to have PE licensure, even though they are the ones who teach and prepare others who take the two required examinations.

PE licensure is thus quite different from board certification for a physician or bar admission for a lawyer. Physicians cannot practice medicine without board certification, and lawyers can't practice law if not admitted to the bar, but such strict limitations do not apply universally to engineers. While the title "engineer" is legally protected in many states (meaning that it is unlawful to use it to offer engineering services to the public without a PE license), there are numerous exceptions. People licensed as "operating engineers" are not PEs. There is an industrial exemption for "in house" engineers employed by a manufacturing company or other business not providing a service directly to the public. Employees of state or federal agencies may also call themselves engineers if that term appears in their official job title.

Courts generally have not required an expert to be a licensed PE in order to testify.³⁹ In some cases, however, PE licensure can be an important factor in determining if an expert is qualified.⁴⁰ In still other cases, courts have held that

37. See Eligibility Requirements for the PE Exam by State, available at <https://perma.cc/D9RV-XBNZ>.

38. Because they often are involved in public works projects, civil engineers are more likely than engineers in other fields to obtain PE licenses. This can be traced directly back to the legacy of the St. Francis dam collapse in southern California in the late 1920s. St. Francis Dam Disaster Revisited (D.B. Nunis, Jr., ed., 2002).

39. See, e.g., *Emig v. Electrolux Home Prods. Inc.*, No. 06-CV-4791 (KMK), 2008 WL 4200988 (S.D.N.Y. Sept. 11, 2008) (expert qualified to testify even though not a PE). Cf. *McRunnel v. Batco Mfg.*, 917 F. Supp. 2d 946 (D. Minn. 2013) (fact that expert licensed in Arkansas but not Minnesota was proffered in Minnesota did not make him unqualified).

40. See, e.g., *Clena Investments, Inc. v. XL Specialty Ins. Co.*, 280 F.R.D. 653, 661–62 (S.D. Fla. 2012) (Citing education and experience of plaintiff's expert and noting that like defendant's expert, he was a professional engineer. "Presumably [defendant viewed its expert's] professional engineering license as at least one relevant qualification . . . No less is true of [plaintiff's expert's]

a PE without adequate experience relevant to the issues at hand is not qualified.⁴¹

Just as an undue judicial gatekeeping focus on an expert's degree or licensure is inappropriate, gatekeeping based only on determining whether scientific or engineering expertise is required would be a futile endeavor. Rather, the inquiry ought to focus on the extent to which the *scientific* hypothesis testing or the *engineering* design process is at issue. As an example, fire investigators are often considered engineers, and some universities even offer courses in fire protection engineering. Yet the conduct of an investigation into the origin or cause of a fire involves the kind of hypothesis testing so essential to science. The National Fire Protection Association's Standard 921 "sets the bar for scientific-based investigation and analysis of fire and explosion incidents."⁴² Note, however, that while the *Daubert* hypothesis-testing factor is important for such investigations, peer review of every investigation would not be appropriate, nor would the general acceptance of every investigation.

By way of contrast, the investigation into the 1981 collapse of a hotel skywalk in Kansas City, which killed more than a hundred people,⁴³ provides a classic example of determination of cause through analysis of a design. Without any testing or peer-reviewed publication, and based only on a careful analytical review (using accepted methods and principles), engineers from the National Bureau of Standards quickly determined that the original design was safe, but that a modification made during construction had resulted in an unintended doubling of the load on one element of the support structure.⁴⁴

comparable license and experience."); *Urda v. Valmont Indus., Inc.*, 561 F. Supp. 3d 632 (M.D. La. 2021) (PE license one factor considered in finding expert qualified); *Klingenberg v. Vulcan Ladder USA, LLC*, 936 F.3d 824 (8th Cir. 2019) (PE license one factor considered in finding expert qualified).

41. See, e.g., *Dreyer v. Ryder Automotive Carrier Grp., Inc.*, 367 F. Supp. 2d 413, 426–27 (W.D.N.Y. 2005) (Although proffered expert held a doctorate in mechanical engineering and was a licensed professional engineer, he had no direct experience with the trailer securing system at issue in the case. When asked to explain how a designer would go about assessing the foreseeability of user risk in connection with a product, he could not answer clearly.); *Taber v. Allied Waste Sys., Inc.*, 642 Fed. App'x 801 (10th Cir. 2016) (PE excluded because of lack of relevant experience).

42. NFPA 921, available at <https://perma.cc/R9RU-MZMF>. See also *Elosu v. Middlefork Ranch Inc.*, 26 F.4th 1017, 1029 (2022) (citing NFPA 921 and admitting investigator's testimony after noting that he "applied broadly accepted scientific principles and professional standards to conduct his analysis").

43. *In re Fed. Skywalk Cases*, 680 F.2d 1175, 1177 (8th Cir. 1982).

44. Deborah R. Hensler & Mark A. Peterson, *Understanding Mass Personal Injury Litigation: A Socio-Legal Analysis*, Brook. L. Rev. 961, 972–74 (1993).

How Do Engineers Think, Make, Test, and Evaluate?

In his book *Applied Minds: How Engineers Think*,⁴⁵ Guru Madhavan explores the mental tools of engineers that allow engineering achievements. He builds his framework around a flexible intellectual tool kit he terms modular systems thinking. We focus on this framework to help explain how engineers think, and supplement our discussion with ideas from the book *The Nature of Technology*⁴⁶ by W. Brian Arthur.

Madhavan's thesis boils down to how engineers think in terms of systems, which means that an engineer can deconstruct (break down a larger system into its modules) and reconstruct (put it back together). The focus is on identifying the strong and weak links—how the modules work, don't work, or could potentially work—and applying this knowledge to engineer useful outcomes. Madhavan contends that there is no engineering method that is broadly applicable across all areas of engineering, so modular systems thinking varies with context. As an example, he describes how engineering the Burj Khalifa (the tallest building in the world, in Dubai) is different, say, from coding the Microsoft Office Suite. Whether used to conduct wind tunnel tests on World Cup soccer balls or to create a missile capable of hitting another missile midflight, engineering works in various ways. Techniques can differ even within a specific industry, but Madhavan identifies the three essential properties of the engineering mindset: the ability to see a structure where there's nothing apparent, adeptness at designing under constraints, and understanding tradeoffs.

Structured Systems-Level Models and Thinking

Our world relies on structure, and the engineering mind gravitates to the unseen portion of the structural iceberg underneath the water rather than what is seen on its surface. A good engineer can visualize and produce structures through a combination of rules, models, and experience. For engineers, the word “model” usually means a mathematical representation of “real world” phenomena. Engineers routinely use mathematical models to represent and conceptually manipulate what happens in the physical world. Once a mathematical model is established and validated, then it can be used to rapidly predict the outcome of an event from a set of inputs.⁴⁷

45. Guru Madhavan, *Applied Minds: How Engineers Think* (2016).

46. W. Brian Arthur, *The Nature of Technology: What It Is and How It Evolves* (2009).

47. Courts have addressed mathematical models in various contexts and generally admit evidence based on them so long as the inputs reflect the actual facts in a case. See, e.g., *Lapsley v. Xtec*,

A structured systems-level thinking process would consider how the elements of the system are linked in logic, in time, in sequence, and in function. Arthur makes a similar point, by describing technologies (i.e., engineering products) as constructions, or “combinations of components and assemblies.”⁴⁸ In addition, an engineer would consider under what conditions the constructions work and don’t work. An historian might apply this sort of structural logic decades after something has occurred, but an engineer has to do it preemptively, whether with the finest details or top-level abstractions. This need to think about what does not yet exist is one of the main reasons engineers often use mathematical models (and sometimes build actual physical models), which facilitate conversations based in reality. Critically, envisioning a structure also involves having the wisdom to decide when it is valuable, and when it isn’t—which is an essential engineering decision.

Adeptness at Designing Under Constraints

Madhavan expounds that, given the practical nature of engineering, the pressures on engineers are greater compared to the pressures experienced by many other professions. Constraints, whether natural or man-made, don’t allow engineers to wait until all matters are fully understood and explained. Engineers are expected to produce the best possible results under the given conditions. Even if there are no fundamental constraints, good engineers know how to apply practical solutions to help achieve their goals. Time constraints on engineers fuel creativity and resourcefulness. Financial constraints and physical constraints imposed by the laws of nature are also common, coupled with an unpredictable constraint—human behavior.

Moreover, because engineers can come up with different (sometimes very different) designs to achieve a specific goal, there is not a single “right” design that fits within given constraints. Proper application of the engineering design process can lead to different end points. For example, engineers have created a wide variety of products to make coffee. As described by Arthur, engineers usually engage in the “planning, testing, and assembly of a new instance of a

Inc., 689 F.3d 802, 816 (7th Cir. 2012) (admitting testimony based on a “mathematical model . . . used in place of physical re-creation”); Leibfried v. Caterpillar, Inc., Case No. 20-CV-1874, 2023 WL 7284795 at *4 (E.D. Wis. Nov. 3, 2023) (expert’s manipulation of “the virtual model to produce results that approximated the post-accident condition” did not preclude admissibility of his testimony); Sommerville v. Union Carbide Corp., Civ. Action No. 2:19-cv-00878, 2024 WL 1204094 at *1 (S.D. W. Va. Mar. 20, 2024) (testimony based on model excluded because the inputs were “speculative and . . . premised on assumptions that do not accurately represent the Defendants’ operations”).

48. Arthur, *supra* note 46.

known technology,” and not necessarily the invention of new technology.⁴⁹ Perhaps the utilitarian notion of providing “the greatest good for the greatest number” is apt here.

Understanding Tradeoffs

Engineering has also been described as a “set of tradeoffs,” and Madhavan introduces the notion of tradeoffs as a last pillar of engineering thought. Tradeoffs are how engineers adapt to constraints. Engineers adopt design priorities and allocate resources by eliminating less important goals and adapting to the more critical constraints in order to focus on achieving the more important goals. When working on the design for a product, engineers must therefore balance various design constraints such as usefulness, usability, cost, performance, aesthetics, and time limitation, all of which make the engineering task quite challenging. A tension exists between the various specifications of the customer. Famed bicycle parts engineer Keith Bontrager is well known for his aphorism: “Strong. Light. Cheap. Pick Two.”⁵⁰ In other words, customers can choose two of three characteristics: strong, light, and cheap. He specialized in designing strong, light, and expensive components. That kind of unavoidable tradeoff leads engineers to rely on their education and training, as well as on their experience and standards to achieve a successful outcome. If one or more of these constraints is relaxed or eliminated (e.g., unlimited time), then the engineering task becomes easier and more achievable, unless the constraint is a fundamental physical law. As another example, for a given airplane design, a typical tradeoff may be balancing the demands of cost, weight, wingspan, and lavatory dimensions within the constraints of the given flight performance specifications. This selection pressure may even lead to the question of whether passengers actually like the airplane they’re flying in and enjoy the experience. If designing within constraints can be viewed as walking a tightrope, then tradeoffs are akin to a tug-of-war among what’s available, what’s possible, what’s desirable, and what the fundamental physical limits are.

Thus, Madhavan’s three essential properties⁵¹ of the engineering mindset lead to the concept of “good engineering practice” or “best engineering practice,” which guides the practitioner of the art.

Arthur makes a different point as he describes engineering design as a matter of choice. As constraints limit possibilities, the design problem becomes more complicated, which leads to a need for a larger number of and more sophisticated components in order to achieve the design goals. Somewhat counterintuitively,

49. *Id.*

50. Grace on Wheels, “Strong, Light, Cheap. Pick Two,” available at <https://perma.cc/BE7H-DPAA> (accessed July 17, 2024).

51. Madhavan, *supra* note 45.

as long as they do not completely eliminate feasibility, constraints actually lead to a richer set of practical possibilities. In effect, “[a]ny new version of a technology is potentially the source of a vast number of different configurations.”⁵²

Engineering Systems

System analysis has become increasingly important in engineering design. Technologies must be evaluated in ever-broadening contexts that include their performance as intended by the designer, but also how they interact with other systems. An electric car, for example, has many mechanical and electrical components as well as software modules that are tested individually, then as subsystems, and finally as a whole system as described in the next section. The hierarchy continues, however, as the car is operating along with other cars over a transportation system. The safety and risk analyses must therefore be conducted at all levels of the hierarchy, from a failure analysis of components all the way up to avoiding accidents in a dynamic traffic environment. The analysis is made more difficult as more products are engineered, monitored, and controlled using machine learning and artificial intelligence.

Systems and Systems of Systems

As mentioned earlier, engineering is a human activity, and while some animals make tools and come together in groups to achieve simple engineering feats, humans are unique in their ability to design systems made of specifically designed components, which may then be assembled into larger systems and systems of systems. A distinguishing feature of engineering is that its products and technologies are intended for use by experts and non-experts alike and must therefore be useful *and* safe as they interact with living and nonliving objects. Understanding and/or operating sophisticated systems or high-complexity systems, such as flying an airplane, will of course require advanced training and certification, while driving a car requires much less. In addition, engineering systems interact with other systems, leading to systems of systems where the interactions between and among systems are just as critical as the behavior of each individual system.

Consider, for example, the social goal of transporting people and goods efficiently while keeping vehicular congestion at an acceptable level. A modern transportation system is composed of vehicles, road networks, railroads, shipping lanes, air traffic lanes, policies, governance structures, and more. A vehicle is one such system, composed of many electrical and mechanical subsystems.

52. Arthur, *supra* note 46.

Putting all the transportation system's subsystems together requires the expertise and contributions of civil engineers, mechanical engineers, electrical engineers, and computer engineers, as well as materials engineers, policy makers, psychologists, and financial and other experts.

Some courts have recognized that engineering products are a collection of components and systems, and the value of expertise in designing and analyzing such systems. *Campbell v. Fawber*⁵³ provides an outstanding example, and shows how engineers can address complex causation questions in the context of litigation. The plaintiff was a passenger in a sport utility vehicle (SUV) that rolled over when the driver lost control.⁵⁴ The plaintiff suffered catastrophic neck injuries that left her quadriplegic.⁵⁵ She sued both the driver and the manufacturer of the SUV, the latter on the theory that the vehicle was not crashworthy because the roof crushed too easily.⁵⁶ She introduced testimony from four experts—three engineers and a specialist in epidemiology and biostatistics. Making this testimony mesh was essentially a matter of systems analysis, and one of the engineering experts explicitly “employed standard scientific systems analysis techniques . . . [which took into account] all aspects of the automotive system through the full crash sequence.”⁵⁷

Complex Systems

The increasing significance of a systems approach to engineering reflects increasing complexity and interaction between products and product components. An article by Robert Lucky cites Samuel Arbesman to explain that there are three main reasons for increasing complexity: accretion, interconnection, and edge cases. Accretion, Lucky writes, “is the result of large systems being built on top of smaller and older systems, often via the incorporation of legacy code, producing what we call kludges.”⁵⁸ He further explains that when “these subsystems become interconnected, the resulting entanglement can change what was simply intricate to truly *complex*.”⁵⁹ “Complexity,” he writes, “is exacerbated by the inevitable existence of edge cases—that myriad of individually negligible exceptions and rarities that yet constitute the long tail of cases that must all be accounted for in system design. The complexity resulting from these factors has passed a tipping point where no single person can fully understand a

53. 975 F. Supp. 2d 485 (M.D. Pa. 2013).

54. *Id.* at 488–89.

55. *Id.* at 489.

56. *Id.* at 500.

57. *Id.* at 492. See *infra* note 158 and related discussion.

58. Robert W. Lucky, *Should Engineers Start Thinking Like Field Biologists?*, IEEE Spectrum, Dec. 20, 2016, available at <https://perma.cc/UN6V-4VYJ>.

59. *Id.*

complete system.”⁶⁰ Complexity may also result from hidden causes, as well as from intentionally introduced design parameters.

Complexity also leads to a less understood phenomenon, namely emergent behavior. As discussed by Anderson,⁶¹ when many objects interact within a system, their overall behavior can no longer be understood using a reductionist approach. As the number of engineered devices multiply, such as within the Internet of Things (IoT), no amount of advanced modeling could *a priori* predict the devices’ ultimate performance. In such a case, there is no substitute for monitoring the actual system in its operational state and designing conservative guardrails. Complexity thus may lead to design flaws that escape detection in the engineering design process.⁶²

Ever since the advent of computing, simulation tools have been used in the design process. While historically, engineering products were designed “by hand,” computers are now ubiquitous as aids to the engineer, and computing tools are used throughout the design and manufacturing processes. The more complicated a system, the more insight may be gained into its behavior by using digital computer simulations. While some systems (such as chaotic systems) may defy causal and deterministic predictions under realistic conditions (slight variations in components), simulations still provide a relatively inexpensive way to model and test engineered components and systems. As more sophisticated tools become available, and with advances in the standard of practice, designers and engineers can model more realistic scenarios, allowing them to push the envelope of performance. Most recently, the availability of more powerful hardware, such as graphical processing units (GPUs), have made machine learning tools more widely available and as such have advanced software development and systems modeling. Below, we discuss two cases where engineering systems failed, and explain the lessons learned.

When Systems Fail

Many cases in which engineers testify arise because an engineered product, structure, or system has failed. Examples of such cases are discussed below in the section titled “Analysis of the Case Law Based on an Understanding of Engineering and How It Differs from Science.” Here we discuss, from an engineering

60. *Id.*

61. P.W. Anderson, *More Is Different: Broken Symmetry and the Nature of the Hierarchical Structure of Science*, 177 *Science* 4047, 393–96 (Aug. 4, 1972), <https://doi.org/10.1126/science.177.4047.393>.

62. Ch. 3–3: *The Role of Failure in Engineering Design: Case Studies*, in *Intro to ME: Design and Analysis* (2003), available at <https://perma.cc/5CE7-3EV4> (see <https://web1.eng.famu.fsu.edu/~chandra/courses/eml3004c/book/printversion/thebook/0130296333.pdf>).

perspective, an aerospace systems failure associated with batteries in a Boeing airplane.⁶³

Among the various engineering challenges Boeing has faced in recent years is the grounding of the Dreamliner fleet because of lithium ion battery fires.⁶⁴

At 10:21 a.m. on [January] 7, 2013, about a minute after all 183 passengers and 11 crew members from Japan Airlines Flight 008 disembarked at Boston's Logan International Airport, a member of the cleaning crew spotted smoke in the aft cabin of the Boeing 787-8. A mechanic then opened the aft electronic equipment bay of the plane, parked at the airport gate, and saw billowing smoke and flames coming from the batteries for the 787's auxiliary power unit (APU).⁶⁵ . . . The culprit was a lithium-ion battery manufactured by GS Yuasa. [The battery] was found to be under a condition known as a thermal runaway, in which the heat from a failing cell causes itself and surrounding cells to fail, thereby generating more heat.

[In December 2014], the U.S. National Transportation Safety Board [NTSB] released its report on the Japan Airlines fire. The agency faulted Boeing, GS Yuasa and FAA for shortcomings in their respective roles in the 787 grounding. [This] episode highlights some of the emerging concerns around cutting-edge clean technologies, particularly those that store large amounts of energy. Saving electrons for later is important for applications like electric vehicles, and denser batteries mean vehicles can go farther with less weight, a major consideration for fuel-conscious airlines. Storing electricity is also crucial for smoothing over peaks and valleys in output from wind and solar [generation facilities]. Batteries, especially the lithium-ion variety, are largely filling this role. But doing so at larger scales and with new systems requires additional scrutiny and training for the unique challenges posed and the growing pains that may arise.

63. For information about Boeing, a global aerospace company that “develops, manufactures and services commercial airplanes, defense products and space systems for customers in more than 150 countries,” see Boeing in Brief, Boeing company website, at <https://perma.cc/CN8R-CGMS>.

64. The description of the Dreamliner batteries incident was reproduced from Umair Irfan, *Climatemire* (published in *Scientific American*, December 18, 2014, available at <https://perma.cc/4CSK-U2GH>).

65. The mechanic used a fire extinguisher to try to put out the flames, “but the blaze didn’t go out. Firefighters arrived at the scene at 10:37 a.m., and using a thermal imaging camera, they looked through the smoke in the equipment bay and saw a softball-sized glow. Responders attempted to extinguish the heat source with Halotron, a fire suppressant. The flames went out, but another firefighter saw that the batteries were still giving off heat and appeared to rekindle. The batteries hissed, leaked fluid and popped. . . . By 12:19 p.m., firefighters had declared the event ‘controlled,’ having removed the APU battery after extinguishing it with further doses of Halotron.” *Id.*

For the Dreamliner incident, the NTSB concluded Boeing did not “incorporate design mitigations to limit the safety effects that could result in such a case.”⁶⁶ Thus, it is evident that system design has to envision failure modes and provide for mitigation solutions that will provide the desired safety level for the system.

What this failure highlighted was the fact that for such complex systems with catastrophic consequences of a failure, the constraints placed on engineers might have led to inadequate solutions. Therefore, engineers testifying in such cases should be allowed to elaborate on the constraints that were placed on their design and on their work in general, as those might have been incompatible with the stated objective.

Computers, Artificial Intelligence, and Machine Learning

The increased use of data in every human endeavor has penetrated all engineering fields, and as a result machine learning (ML) and artificial intelligence (AI) tools have become increasingly important. Large language models (LLMs) (e.g., ChatGPT) have thrust these tools into our daily lives and into the engineering process. LLMs are being used to conduct patent searches and to assist in software writing. While this development has led to performance improvements, it has created new concerns as we try to evaluate engineered systems and to assign responsibility when systems fail to perform as expected. Moreover, systems designed using ML techniques often lack transparency in their internal workings and may become harder to explain and debug. In fact, the National Science Foundation’s Engineering Research Visioning Alliance (ERVA) has come out with a report that advocates creating AI engineering as a new discipline.⁶⁷

Even though these technologies are now seemingly ubiquitous, we shouldn’t overlook how truly revolutionary they are and the remarkable things they enable us to do today and will allow us to do tomorrow. For engineers, AI and ML might cause the tasks they perform to evolve, but these tools can also help them perform new tasks they couldn’t undertake before. There likely will be unintended and unforeseen consequences, however, as use of the new technologies becomes ever more common. As discussed in the *Reference Guide on Artificial Intelligence*, in this manual,⁶⁸ unwitting human bias “refers to the unintentional infusion into an application of human preferences, stereotypes, values, fears, or knowledge. Consider an algorithm intended to predict risk. An engineer might apply engineering

66. Boeing, *supra* note 64.

67. AI Engineering A Strategic Research Framework to Benefit Society, Engineering Research Visioning Alliance, 2024, accessible at <https://perma.cc/9377-WMDW>.

68. See James E. Baker & Laurie N. Hobart, *Reference Guide on Artificial Intelligence*, p. 1525, in this manual.

principles to a risk equation.” An excellent example of unwitting human bias is described in a study published in *Science*⁶⁹ in 2019. In that study, it was found that an algorithm widely used in U.S. hospitals to allocate health care to patients has been systematically discriminating against black people—the algorithm was less likely to refer black people than white people who were equally sick to programs that aim to improve care for patients with complex medical needs. The study states that hospitals and insurers use the algorithm and others like it to help manage care for about 200 million people in the United States each year. Engineers may unwittingly succumb to other unknown biases because ML algorithms are based on training data generated and selected by humans. Bias in the training data will lead to biased algorithms. For more detailed discussion of the use of AI in engineering and elsewhere, please see the *Reference Guide on Artificial Intelligence*,⁷⁰ in this manual.

Analysis of the Case Law Based on an Understanding of Engineering and How It Differs from Science

To gain insight into the kinds of cases in which engineering experts testify and the issues they address, we searched the Westlaw “All Federal” database for all cases decided over the 18 months from January 1, 2021, through June 30, 2022, in which “*Daubert*,” “*Kumho*,” or “702” appeared, along with “engineer” or “engineering” in the same sentence as “expert.” Of the 496 cases we found, only 201 (i.e., about 134 cases a year) required a court to address the question of what constitutes valid and reliable engineering testimony.⁷¹ An explanation of how we winnowed down to 201 cases is provided in the footnote below.⁷²

69. Ziad Obermeyer et al., *Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations*, *Science* 366, 447–53 (2019), <https://doi.org/10.1126/science.aax2342>.

70. See James E. Baker & Laurie N. Hobart, *Reference Guide on Artificial Intelligence*, section titled “Bias,” in this manual.

71. Of course, the sample did not include the many cases that settle or otherwise resolve without leaving any trace in the Westlaw system.

72. We first eliminated from the Westlaw sample all cases that were not on point. One hundred thirty-eight of 496 cases did not really relate to engineering at all, or were reversed. Of the remaining 358 cases, 76 involved the kind of simple accident investigations with which courts have a long history, and for which judges do not likely require any reference materials beyond what the parties provide. Taking out the 76 simple accident cases from the remaining 358 Westlaw search results left 282 cases for further consideration. The number was reduced yet further by taking out insurance coverage cases. The engineering issues in such cases are confined almost exclusively to causation and qualifications, and the use of experts goes back a long time. In *Peele v. Merchants’ Ins. Co.*, 19 F. Cas. 98, 102 (D. Mass. 1822), the court noted that with regard to the sufficiency of repairs to an insured ship, it would if necessary use “experts to report upon the whole evidence, what in their judgment is the true posture of the case in this respect.” After we removed the

In the discussion below, we have categorized the 201 relevant cases by the kind of litigation involved, and within each category (based both on the sample cases and further research) we consider how courts have distinguished between engineering and science, and when and how they have focused on some aspect of the design process or the way in which engineers evaluate designs. Products liability claims were at issue in the majority of the 201 cases, but there were a number of contract or construction litigation cases, intellectual property cases, and environmental or land use cases. We also discuss two other categories: cases involving failure analysis (which often overlap with products liability cases), and cases in which some kind of engineering test (e.g., a test for the strength of steel) was at issue. If and how to apply the *Daubert* factors to engineering is a recurrent theme in most of the categories, as exemplified by *AZZ, Inc. v. Southern Nuclear Operating Company, Inc.*,⁷³ in which the court found an engineer's testimony inadmissible as scientific evidence, but admissible because he employed an appropriate experience-based methodology.

Products Liability Cases

According to Professor David Owen, “Americans suffer some 200 million non-trivial product-related injuries and illnesses each year, at an annual cost to the nation approaching \$10 trillion.”⁷⁴ Most of these injuries are the result of product misuse, but “many product-related accidents result from dangers in or misrepresentations about products that manufacturers reasonably should have taken steps to avoid. It is cases involving these types of product deficiencies that are the realm of products liability law.”⁷⁵

In most American legal jurisdictions⁷⁶ there are four legal theories under which a products liability plaintiff can recover: (1) negligence, (2) tortious misrepresentation, (3) breach of warranty, and (4) strict liability.⁷⁷ Putting aside

accident and insurance coverage cases, 226 cases remained. Of these, another 25 involved matters like employment related claims, criminal law issues, or civil rights that similarly don't present difficult engineering evidence admissibility questions. That left 201 cases in which the admissibility challenge was potentially more complex, or about 134 per year. One hundred twenty-six of the 201 cases involved products liability. Twenty-four involved contract or construction litigation claims, 22 involved intellectual property, and 19 involved environmental or land use questions.

73. CV-119-052, 2023 WL 2746033 at *13 (S.D. Ga. Mar. 31, 2023).

74. David G. Owen, *Products Liability in a Nutshell* 12–13 (10th ed. 2022).

75. *Id.* at 13.

76. In 1992, the American Law Institute commissioned two professors “to prepare a new *Restatement (3d) of Torts* on the specific topic of products liability law.” Owen, *Products Liability in a Nutshell*, *supra* note 74, at 19. The new *Restatement* was approved in 1997 and published in 1998, but most jurisdictions have not formally adopted it, and most courts at least nominally still adhere to the traditional products liability legal framework. *Id.* at 19–20.

77. *Id.* at 15–20.

tortious misrepresentation, a relatively rarely used theory, “the plaintiff in nearly every products liability case must prove that a product sold by the defendant contained an unnecessary hazard [or defect] that caused [physical] harm,”⁷⁸ and most engineering expert testimony in products liability cases relates to whether there was or wasn’t a defect, and if there was, whether it caused the plaintiff’s injury.

Modern products liability law dates from the late 1950s and early 1960s. Over the past 60 or so years, “courts and commentators [have come] to understand the distinctions between three very different forms of product defect.”⁷⁹ Manufacturing defects are “unintended physical irregularities that can occur during the production process.”⁸⁰ Design defects are “hazards in a product’s engineering or scientific conception that reasonably should be avoided by a different design or formula.”⁸¹ Warning defects occur when there is an “absence of important information about product hazards or how to avoid them.”⁸² Below, we discuss how engineering expert testimony is evaluated with regard to each kind of defect and with regard to the issue of causation. We also discuss engineering testimony on a manufacturer’s duty of care, an issue that arises when a plaintiff claims under a negligence theory.

Design Defects

Depending on the jurisdiction and the facts in a given case, the legal test for determination of a design defect can turn on either (1) what an ordinary consumer might reasonably expect, or (2) a balancing of a product’s risk versus its utility or benefit.⁸³ Plaintiffs generally may introduce expert testimony to establish a defect under either theory, but an expert will more often be necessary to prove a defective design under the risk–benefit test. See, for example, *Bradley v. Ameristep, Inc.*,⁸⁴ in which the court noted that under Tennessee’s products liability law, a plaintiff invoking the consumer–expectation test can prevail without any expert testimony at all. The consumer–expectation test generally does not apply to complex products, however. As explained in *Brown v. Raymond Corp.*,⁸⁵ consumers have no reasonable expectations about products so complex they can’t really be understood.

78. *Id.* at 20, and *Restatement (Second) of Torts* § 402A.

79. Owen, *supra* note 74, at 199.

80. *Id.*

81. *Id.*

82. *Id.*

83. See Clayton J. Masterman & W. Kip Viscusi, *The Specific Consumer Expectations Test for Product Defects*, 95 Ind. L. J. 183, 184–85 (2020), and Ramirez v. ITW Food Equip. Grp., LLC, 686 Fed. App’x 435, 437 (9th Cir. 2017) (consumer–expectations and risk–benefit tests provide alternative means for a plaintiff to prove a design defect). See also Owen, *supra* note 74, at 186–87.

84. 800 F.3d 205, 211 (6th Cir. 2015).

85. 432 F.3d 640, 643 (6th Cir. 2005).

Given the limited use of experts in cases based on consumer expectations,⁸⁶ we only address risk–benefit cases here. In such cases, the predominant engineering issue is the proper way to develop and evaluate designs. Two of the most significant legal questions are (1) whether an expert has to identify a reasonable alternative design, and (2) whether testing is required to prove the existence of a defect or to validate a proposed alternative design. With regard to the *Daubert* factors, testing is the only factor typically at issue. The peer-review/publication process generally does not apply for designs, whether original or alternatives proposed in litigation. While they are reviewed and checked, designs typically do not go through peer-reviewed publication. Nor is the question of error rate meaningful in the analysis of most designs. Sometimes general acceptance of an alternative design may be considered, as well as design or performance standards, but of the *Daubert* factors, testing is what most often becomes an issue in products liability cases.

Reasonable alternative design

The *Brown v. Raymond Corp.* case is an example of requiring evidence of a safer alternative design. The plaintiff was injured in a forklift accident, which he claimed was caused by a defective design.⁸⁷ The trial court excluded testimony from the plaintiff’s design-defect expert because he “was not an expert in forklift design and had not proposed any alternative designs against which to test his conclusion that the Raymond forklift was unreasonably dangerous due to a design defect.”⁸⁸ The Sixth Circuit affirmed.⁸⁹ Note, however, that the case did not turn on any issue about engineering testimony because the plaintiff had asserted claims based on the consumer-expectation test, and there was no alternative design to evaluate.⁹⁰

Notwithstanding cases like *Brown*, an alternative design is not absolutely necessary even in risk–benefit cases. *Williams v. Tristar Products, Inc.*⁹¹ illustrates this

86. Also, as Professor Owen points out, courts increasingly are turning to the risk-benefit test. David G. Owen, *Products Liability in a Nutshell* 175 (10th ed. 2022).

87. *Brown*, 432 F.3d at 641–42.

88. *Id.* at 647.

89. *Id.* at 650. See also *Lara v. Delta Int’l Mach. Corp.*, 174 F. Supp. 3d 719, 736 (E.D.N.Y. 2016) (“In a products liability action, ‘the “touchstone” of an expert’s report should be a comparison of the utility and cost of the product’s design and alternative designs.’”); *Peck v. Bridgeport Machs., Inc.*, 237 F.3d 614, 618 (6th Cir. 2001) (reasonable alternative design required); *Willet v. Johnson & Johnson*, 465 F. Supp. 3d 895, 904 (S.D. Iowa 2020) (“plaintiffs . . . [must] demonstrate the existence of a reasonable alternative design”).

90. *Brown*, at 643 (plaintiff’s counsel conceded the case depended on applying the consumer-expectation test).

91. 418 F. Supp. 3d 1212 (M.D. Ga. 2019).

point. The plaintiff in the case suffered burns when the lid of a pressure cooker unexpectedly popped off. The court admitted testimony from her expert, who had “a Ph.D. in engineering and . . . extensive experience in design and development of products, fasteners, machines, tools, and latching and locking devices.”⁹² He had tested a pressure cooker like the one that injured the plaintiff, and explained his testing in detail.⁹³ “His familiarity with pressure cookers manufactured by defendant gained by testifying as an expert in similar litigation further demonstrate[d] his experience and knowledge of the product.”⁹⁴ The expert had not proposed any alternative design, but the court held that “it is not required that either plaintiff present evidence of reasonable alternative designs or that the evidence be in the form of expert testimony.”⁹⁵ Though it did not involve a proposed alternative design, *Williams* falls squarely within the kind of analysis suggested in this reference guide because the issue was the expert’s analysis of an *existing* design.⁹⁶

*Is testing required to prove existence of a defect
or the validity of an alternative design?*

The question of testing arises in many products liability cases, both to establish a defect and to validate a reasonable alternative design. The focus on testing usually derives from *Daubert*, not *Kumho Tire*, and while courts sometimes do suggest testing is part of engineering practice as well as science, they also may confuse or miss the distinction between scientific testing of hypotheses and engineering analysis and testing of designs. For example, in *Colon ex rel. Molina v. BIC USA, Inc.*,⁹⁷ the court held that “[w]hile testing is not an ‘absolute prerequisite’ . . . it is usually critical to show that an expert ‘adhere[d] to the same standards of intellectual rigor that are demanded in their professional work.’”⁹⁸ The court then discussed what *scientists* typically do, stating that “[a]dherence to *engineering* standards of intellectual rigor almost always requires testing of a *hypothesis*.”⁹⁹

92. *Id.* at 1222.

93. *Id.*

94. *Id.*

95. *Id.* at 1229.

96. *Cf. Bullock v. Volkswagen Grp. of Am., Inc.*, 107 F. Supp. 3d 1305, 1315 (M.D. Ga. 2015) (court more or less reasoned in reverse to hold that testimony about a reasonable alternative design was sufficient to establish a design defect, even though the plaintiff’s expert was reluctant to describe the problem he saw as a design defect).

97. 199 F. Supp. 2d 53 (S.D.N.Y. 2001).

98. *Id.* at 76 (quoting *Cummins v. Lyle Indus.*, 93 F.3d 362, 369 (7th Cir. 1996)).

99. *Id.* (emphases added).

While the decision may well have been sound, it appears to conflate science and engineering.¹⁰⁰

*Nease v. Ford Motor Co.*¹⁰¹ provides an example of testing to establish if there really was anything at all wrong with the product in question. The plaintiffs claimed that a cable housing design was defective because it allowed dirt and other contaminants to interfere with the operation of a pickup truck accelerator cable. They claimed a crash had resulted when a stuck cable prevented the truck in question from decelerating after the driver let up on the gas pedal. The trial court admitted the testimony of the plaintiffs' expert, an electrical engineer, but the Fourth Circuit reversed, holding that the expert's analysis of the design was inadequate. "Testing was of critical importance in this case . . . [but the expert had] conducted no testing whatsoever to arrive at his opinion. . . . [He had not determined] whether it is actually possible for enough debris to accumulate in the casing cap during normal operation to resist the . . . force exerted by the return springs to pull the throttle closed."¹⁰²

By way of contrast, in *Lapsley v. Xtek, Inc.*,¹⁰³ an expert's opinion on a design defect was admitted without testing. The plaintiff was severely injured when a high-intensity jet of grease spurted from a bearing on a huge steel rolling mill machine. His expert explained how the design of the bearing lubrication system differed from a design used on similar rolling mills manufactured both before and after the accident. The expert also explained how this difference caused grease

100. *Cf. Amica Mut. Ins. Co. v. Electrolux Home Prods., Inc.*, 440 F. Supp. 3d 211, 217–18 (W.D.N.Y. 2020) (referring to quintessentially engineering tests as scientific tests while applying *Daubert* to engineering testimony); *Morasch Meats, Inc. v. Ludwick*, Case No. DK 18–02714, 2019 WL 5616261, at *2 (W.D. Mich. Sept. 11, 2019) (engineering expert's testing "is the staple of scientific method, even if his assistance in this matter will not be, strictly speaking, scientific").

101. 848 F.3d 219 (4th Cir. 2017).

102. *Id.* at 231–32. *See also* *Elkharroubi v. Six Flags Am. LP*, Civ. Action No. TJS-17–2169, 2020 WL 1043304, at *6–7 (D. Md. Mar. 4, 2020) (plaintiff claimed a malfunctioning trap door on a water slide caused his injury, but expert could not "point to a single observation of the trap door mechanism that would render his opinions plausible"); *Kesse v. Ford Motor Co.*, Case No. 14-cv-6265, 2020 WL 832363 (N.D. Ill. Feb. 20, 2020) (theory that electromagnetic interference could cause sudden acceleration was untested); *Graves v. Mazda Motor Corp.*, 675 F. Supp. 2d 1082, 1102 (W.D. Okla. 2009) (In case involving design of automobile gear shift mechanism, expert's conclusion that shifter was unsafe because it was different from other designs was excluded because it was "not grounded in any objective data or specifically applicable engineering standards. Although [his] Rule 26 report recites his substantial experience with engineering testing in a variety of contexts . . . he did no testing to quantify . . . any exceptional propensity of the gated shifter on the Mazda6 to cause driver confusion about the actual position of the shift lever."). *Cf. Alfred v. Caterpillar, Inc.*, 262 F.3d 1083, 1088–89 (10th Cir. 2001) (expert should have been allowed to testify that a rotary dial speed control on a paving machine was not in compliance with SAE standards, but his testimony on whether this non-compliance made the machine defective was properly excluded because he had done nothing more than cite to the standard; court affirmed summary judgment for the defendant).

103. 689 F.3d 802 (7th Cir. 2012).

to spurt out. The trial court admitted his testimony, noting that he had employed “commonly known methodologies and physics calculations,” which “included the use of Bernoulli’s equation regarding energy in moving fluids and reference to ‘widely accepted factors concerning the force necessary to penetrate human skin.’”¹⁰⁴ The Seventh Circuit affirmed, holding that “the math and science here are within the comprehension of judges and lawyers without extraordinary assistance. Xtek disputes the completeness and therefore the relevance of [the expert’s] calculations, but it has not identified, and we have not detected, any grave questions about the reliability of the calculations [he] actually performed.”¹⁰⁵

Sometimes an engineer’s experience alone may suffice for admissibility purposes without a requirement for testing of the product at issue or explaining how the analysis of a design supports an expert’s conclusions. In *Bullock v. Volkswagen Group of America, Inc.*,¹⁰⁶ the plaintiffs claimed a defective turbocharger on a Volkswagen Passat had caused the vehicle to crash when it accelerated uncontrollably. The court denied the defendant’s *Daubert* motions and allowed a mechanic to testify that a leaking seal had caused the unintended acceleration. He had “investigated and diagnosed unintended acceleration events in cars, and [had] designed a solution to fix a problem with unintended acceleration in prior model Volkswagen turbocharged engines.”¹⁰⁷ He “formed his opinions by observing the physical evidence, testing an exemplar vehicle, and drawing on his extensive experience in the field.”¹⁰⁸ Another expert, a materials engineer, was allowed to

104. *Id.* at 807–08.

105. *Id.* at 809–10. See also *Russell v. Howmedica Osteonics Corp.*, No. C06-4078-MWB, 2008 WL 913320, at *5–7 (N.D. Iowa Apr. 2, 2008) (on the question of whether a certain titanium alloy was properly used in a spinal fixator, plaintiff’s expert was allowed to testify based on articles about the relative merits of various metals, even though there was little evidence about when one metal should be used instead of another, or about her opinion that the alloy in question should not be used in patients weighing more than 180 pounds); *Dunton v. Arctic Cat, Inc.*, 518 F. Supp. 2d 296, 300 (D. Me. 2007) (mechanical engineer and a product designer who worked for defendant were allowed to testify about the adequacy and safety of design for snowmobile steering and suspension components); *Freitas v. Michelin Tire Corp.*, No. CIV. 3:94CV1812 (DJS), 2000 WL 424187, at *2 (D. Conn. Mar. 2, 2000) (expert on tire design defect allowed to testify based on “mathematical calculations and experience conducting burst tests on mismatched tires, as well as upon his review of burst tests performed by other reliable testers, sworn testimony and reports of other experts, and industry and tire company documents”). Cf. *Shreve v. Sears, Roebuck & Co.*, 166 F. Supp. 2d 378, 424 (D. Md. 2001) (expert testimony excluded, but “under the unique circumstances of this case, a lay jury could reasonably conclude without the aid of expert testimony that the snow thrower purchased by [plaintiff] was defective [in design] and unreasonably dangerous”); *Valente v. Textron, Inc.*, 932 F. Supp. 2d 409, 426–27 (E.D.N.Y. 2013) (computer simulation not reliable because expert used the wrong coefficient of friction).

106. 107 F. Supp. 3d 1305 (M.D. Ga. 2015).

107. *Id.* at 1310.

108. *Id.* at 1311.

testify based only on his experience with automotive component analysis and aviation turbochargers that there were safer seal designs.¹⁰⁹

The three lines of cases: *Nease* (testing required to prove defect); *Lapsley* (explanation of analysis is sufficient); and *Bullock* (reliance on expert's experience is sufficient) are not inconsistent. All three relate to the question of what caused an unwanted event, and the different legal analyses reflect the fact that different kinds of experts in differing cases can address the question in different ways. Sometimes the answer is evident to a person with adequate experience, and no testing or analysis is required. In such cases, neither the engineering design process nor the scientific method would be implicated. Sometimes the expert has essentially formulated a hypothesis, and admissibility properly would require *Daubert*-like testing against observational or experimental data. And sometimes an engineering analysis will suffice if it properly reflects the way engineers review and analyze designs. Thus, before deciding on the admissibility of expert testimony, a court needs to first identify what approach the expert has taken, and then apply the appropriate legal analysis.

For plaintiffs in products liability cases, a requirement that an expert must test an alternative (likely by having to build a prototype) can be difficult if their expert has come up with a new design rather than citing to another existing product. In *Lara v. Delta International Machinery Corp.*,¹¹⁰ the expert testified that a table saw without an interlocking safety device was inherently dangerous, but he conceded that “other than some theoretical musings made during his deposition as to how such a design could be accomplished . . . he [had] never actually designed such an interlock system for use in a table saw.” Nor had he done any testing, which the court held should be part of the “utility versus cost comparison.”¹¹¹ Thus, the testimony was excluded.¹¹²

109. *Id.* at 1313. *See also* *Great N. Ins. Co. v. BMW of N. Am. LLC*, 84 F. Supp. 3d 630 (S.D. Ohio 2015) (engineer with a “range of experiences in vehicle mechanics and design, including some experiential background in aerodynamics, and extensive experience in assessing car defects as it relates to fire defects, and prevention” was allowed to testify on whether design defect was implicated in causing an automobile fire); *Sloan Valve Co. v. Zurn Indus., Inc.*, No. 10–cv–00204, 2013 WL 4052030, at *5 (N.D. Ill. Aug. 12, 2013) (“[Engineering expert] applies his previous education and vast experience to give his expert opinion on whether [defendant’s product] was adequate from a mechanical engineering point of view, thereby aiding the fact-finder in determining the ultimate question of willfulness.”); *Bowersfield v. Suzuki Motor Corp.*, 151 F. Supp. 2d 625 (E.D. Pa. 2001) (plaintiff riding in the rear cargo area of a Suzuki Samurai vehicle was injured when it was hit by another vehicle. Based on experience designing occupant protection systems, plaintiff’s expert was allowed to testify that the Samurai design should have included a seat with a seatbelt in the cargo area or a barrier to keep passengers from sitting there; but expert was not allowed to testify about sufficiency of warnings.).

110. 174 F. Supp. 3d 719, 736 (E.D.N.Y. 2016).

111. *Id.*

112. *Id.* at 738. *See also* *Colon ex rel. Molina v. BIC USA, Inc.*, 199 F. Supp. 2d 53, 75–78 (S.D.N.Y. 2001) (requiring testing of alternative design); *Wright v. Case Corp.*, 2006 WL 278384,

Not all courts have required testing of an alternative design, however, especially when it is clearly explained or related to existing designs by analogy. In *Lynn v. Yamaha Golf-Car Co.*,¹¹³ for example, testimony about an alternative design for passenger restraint features on a golf cart was held to be admissible, even though the expert had conducted no testing. He cited other golf carts that included features similar to what he proposed,¹¹⁴ and his opinions were not based on “alternative designs that would require enhanced technology or a complete overhaul of the hip restraint.” While they might “impede on a passenger’s ability to enter and exit the vehicle with ease, a jury could readily find that such an imposition would be insignificant compared to the increased safety users would enjoy by the alternative design.”¹¹⁵ Likewise, in *Cruz v. Kumho Tire Co., Inc.*,¹¹⁶ the expert had not physically tested her proposed alternative design, and could not identify any manufacturer that used it, but her testimony was admissible as she “utilized accepted engineering calculations and risk-utility analysis in developing her design, which draws support from relevant industry studies.”¹¹⁷

Unlike the “did a defect cause the event” cases discussed above, the “is the alternative reasonable” cases almost all require consideration of how engineers

at *4 (N.D. Ga. Feb. 1, 2006) (expert conceded he had done no testing of alternative design, and court held that “testing is particularly important in alternative design cases . . . where it is necessary to show that an alternative design is actually feasible”); *Peck v. Bridgeport Mach., Inc.*, 237 F.3d 614, 618 (6th Cir. 2002) (testimony inadequate where expert said he would have designed a lathe differently, with a “safer” lever shift mechanism, but “had never fabricated such a design and, in fact, had never seen his proposed design used on a lathe anywhere”). *Cf. Crawford v. ITW Food Equip. Grp., LLC*, 977 F.3d 1331, 1340 (11th Cir. 2020) (expert testimony admissible where he built and demonstrated the alternative design he proposed for a meat saw).

113. 894 F. Supp. 2d 606 (W.D. Pa. 2012).

114. *Id.* at 630.

115. *Id.* at 632.

116. Nos. 8:10–cv–219 (MAD/CFH), 8:12–cv–200 (MAD)/(CFH), 2015 WL 2193796 (N.D.N.Y. May 11, 2015).

117. *Id.* at *11. *See also* *Schenone v. Zimmer Holdings, Inc.*, No. 3:12–cv–1046–J–39MCR, 2014 WL 9879924, at *8 (M.D. Fla. July 30, 2014) (Expert allowed to testify about alternative design for artificial hip based on his education and experience, which “underpinned his opinion that [defendant’s acquisition of designs for similar products] provided [defendant] with an available and reasonable alternative design [for the product at issue].”); *Thierfelder v. Virco, Inc.*, 502 F. Supp. 2d 1025, 1030–31 (W.D. Mo. 2007) (admitting opinions of expert with years of experience “in the design of wheels and the dimensions of wheelbases for various kinds of mobile furniture” in case involving injury caused by a cart used for folding chairs; expert did no computer modeling or calculations, but did observe the cart and explained proposed changes to protect the feet of users). *Cf. Thomas v. FCA US LLC*, 242 F. Supp. 3d 819, 831 (S.D. Iowa 2017) (in a case involving allegedly defective airbag, experts did not have to design an alternative airbag because, in their view, defendant already had viable alternatives); *Nicholson v. Biomet, Inc.*, 537 F. Supp. 3d 990, 1010 (N.D. Iowa 2021) (existing products could serve as reasonable alternative design); *Borsack v. Ford Motor Co.*, No. 04 Civ. 3255(PAC), 2007 WL 2142070, at *9 (S.D.N.Y. July 26, 2007) (expert opined there were three alternative door latch designs without citing any testing, though the court did not address whether they were all already in use).

actually go about creating and evaluating designs. Based on that standard, testing is not an absolute requirement, nor is it necessary in many cases. Prototypes are not always required before executing a design, and while moving from design to manufacture of a product may require testing of certain components, that does not mean establishing reasonable feasibility requires such testing. The result in the *Lara* case (exclusion of expert testimony on alternative design for a power saw) might well have been different had the expert fully explained how the proposed alternative design could feasibly have been executed.

Manufacturing Defects

In manufacturing defect cases, the problem with the product is often quite evident, and close gatekeeping scrutiny is not required. For example, in *Schmude v. Tricam Industries, Inc.*,¹¹⁸ the parties agreed a ladder was defective because of an improperly installed rivet, but the defendant nonetheless challenged the plaintiff's expert based on a lack of testing and peer review. The court held that the "defect, by its very nature, did not lend itself to the kind of empirical testing *Daubert* envisioned for opinions that are based on cutting edge science. . . . Once [the ladder] failed, there was nothing to test. . . . Other than carefully examining the failed rivet, including the portion embedded in the backing plate, there seemed little else for an expert to do."¹¹⁹

Other cases have subjected engineering experts to closer scrutiny, especially when testing is reasonably possible and would provide important information. In *Myrick v. U.S. Saws Inc.*,¹²⁰ the plaintiff was injured when a power saw blade manufactured by the defendant broke apart. The plaintiff's expert attempted to eliminate all possible causes except for a manufacturing defect, and eventually concluded the failure had occurred because of a defective weld. The court rejected this approach,

118. 550 F. Supp. 2d 846 (E.D. Wis. 2008).

119. *Id.* at 852. See also *Martin v. Apex Tool Grp., LLC*, 961 F. Supp. 2d 954, 960–61 (N.D. Iowa 2013) (expert's opinion that flaws in pry bar caused it to fail held admissible based on detailed examination of the bar—which led him to conclude failure originated at location of flaw—size of flaw, location of flaw, the way bar was used, and absence of evidence of misuse; court cited *Kumho Tire* for proposition that "an expert might draw a conclusion from a set of observations based on extensive and specialized experience"); *McCloud v. Terex Telelect Inc.*, No. Civ–08–433–D, 2010 WL 3259823, at *3 (W.D. Okla. Aug. 17, 2010) (expert engineering testimony on manufacturing defect in step on rear of a truck admitted because expert "testified in his deposition . . . that he employed a recognized method of engineering analysis in reaching his conclusions"); *Engler v. MTD Prods., Inc.*, No. 13–CV–575 (CFH), 2015 WL 900126 at *2 (N.D.N.Y. Mar. 2, 2015) (plaintiff was thrown off a virtually new riding mower after the brakes failed because of excessive wear; engineer's testimony—that brakes likely had not been adjusted right when the mower was sold—was admitted even though he did not know what had caused the brakes to come out of adjustment or why they were prematurely worn).

120. No. C11–1837Z, 2013 WL 766192 (W.D. Wash. Feb. 28, 2013).

holding that the expert's testimony was not reliable. Expert "was aware of multiple destructive and non-destructive tests available that could have been done in this case, but [he] conducted no tests on the weld or blade at issue."¹²¹

*Pugh v. Louisville Ladder, Inc.*¹²² provides an excellent example of how courts sometimes consider both empirical testing and analytical explanation in assessing engineering testimony on a manufacturing defect. The plaintiff, injured when he fell from a ladder, attributed the accident to manufacturing defects that caused the ladder to fail. In particular, he claimed "microscopic cracks at the ladder's rivets . . . [had] propagated into larger cracks causing catastrophic failure/buckling that resulted in [his] fall."¹²³ The defendant ladder manufacturer maintained that the buckling occurred when the plaintiff slipped and fell on top of the ladder.¹²⁴ In other words, the case came down to whether cracks and buckling caused the fall or whether the fall caused the buckling. The jury returned a plaintiff's verdict, and the defendant appealed, arguing, among other things, that the trial court had abused its discretion in admitting testimony from the plaintiff's engineering experts. The Fourth Circuit affirmed after rejecting the defendant's claim that "principles of physics would disprove [plaintiff's] experts' theory."¹²⁵ The plaintiff's experts initially had concluded the ladder failed based solely on a visual inspection, but thereafter they "performed several tests to support their initial assessment."¹²⁶ Based on this testing and their experience, they "determined that their structural failure theory was scientifically supported by the facts of this case and the most likely cause of the accident."¹²⁷ They also performed "testing and analysis to disprove the opposing theory [of] impact damage."¹²⁸

In many cases, design defect and manufacturing defect claims are alleged together, and the distinction sometimes gets blurred in the admissibility analysis. In *Tillman v. C.R. Bard, Inc.*,¹²⁹ for example, the plaintiff claimed that a defective filter had been implanted in one of her veins to keep blood clots from migrating

121. *Id.* at *5. See also *Easterling v. Ford Motor Co.*, 303 F. Supp. 3d 1211, 1222–23 (N.D. Ala. 2018). Plaintiff claimed a seat belt had come unlatched in the course of an accident as a result of a manufacturing defect. His expert engineer's testimony was excluded because expert admitted he did not know why the buckle had come undone, nor had he done any testing.

122. 361 F. App'x 448 (4th Cir. 2010).

123. *Id.* at 450.

124. *Id.*

125. *Id.* at 453.

126. *Id.* at 455.

127. *Id.*

128. *Id.* While very well reasoned, the court's opinion did conflate engineering with science, probably because that is how the parties argued the case. For example, the court noted that the plaintiff's experts "determined that their structural failure theory was scientifically supported by the facts of this case and the most likely cause of the accident." 361 Fed. App'x at 455. The court cited *Kumho Tire* once and *Daubert* 24 times.

129. 96 F. Supp. 3d 1307 (M.D. Fla. 2015).

to where they could cause serious harm. Her doctor planned to remove the device when her clotting risk subsided, but could not do so because it had tilted and become stuck in the vein. Although it had not failed, she claimed that owing to design and/or manufacturing defects it was prone to failure, and that she'd suffered harm because she would be forced to endure monitoring to avoid a failure-related injury. The plaintiff proffered five experts, ranging from an epidemiologist to several engineers, and while the court excluded some of them, all the engineering testimony on manufacturing defects was admitted, even though none of the experts could examine or test the plaintiff's filter because it remained implanted in her vein.¹³⁰ Whether the design or a manufacturing defect caused this conundrum was not really addressed.¹³¹ The court ultimately denied a motion for summary judgment on the design and manufacturing defect claims simply "because there are issues of fact as to whether the filter's risks are unreasonable as a result of its arguably defective design or manufacture."¹³²

Just as courts sometimes do not require distinctly different testimony on design versus manufacturing defect, they also sometimes do not require an expert to specify the exact nature of the defect that caused an injury. In *Rudd v. General Motors Corp.*,¹³³ the plaintiff was injured while tuning a pickup truck engine. As he leaned over the motor to turn the distributor housing, a blade came off the fan and struck him.¹³⁴ His expert could not pinpoint the cause of the failure, other than to say he thought "the fan blade had to have some small defect for the fatigue failure to start."¹³⁵ He said that, based on what he had encountered in his experience, there were three possible explanations, but also said there could be others.¹³⁶ He could rule out catastrophic or abnormal use, however, as well as the development of lesser damage that might have occurred post-manufacture.¹³⁷ "Although . . . he did not find direct evidence of a manufacturing defect, such as a microscopic grind mark or inclusion near the fatigue-fracture origin, his testimony [was] replete with circumstantial evidence that—through a process of eliminating alternative explanations—might support a finding of a manufacturing defect."¹³⁸

130. *Id.* at 1319–22.

131. *Id.* at 1354.

132. *Id.* See also *Siegel v. Dynamic Cooking Sys., Inc.*, 501 F. App'x 397 (6th Cir. 2012) (plaintiff need not establish whether product had either a design or a manufacturing defect).

133. 127 F. Supp. 2d 1330 (M.D. Ala. 2001).

134. *Id.* at 1340.

135. *Id.*

136. *Id.* at 1341.

137. *Id.*

138. *Id.* at 1342.

Warning Defects

Experts on warnings often have engineering degrees, but their expertise more typically relates to fields like human factors, ergonomics, or communications. Although both human factors and ergonomics might be considered engineering specialties, separate experts are often used for warning defects.¹³⁹ Importantly, the issue of admissibility rarely if ever implicates the creation or analysis of designs, which is the focus of this reference guide. Indeed, some courts have held that no expert testimony is required on the adequacy of a warning.¹⁴⁰

That said, some courts do allow warnings testimony from design experts, especially when the expert has experience with warning issues. In *Thierfelder v. Virco, Inc.*,¹⁴¹ the plaintiff was injured when a cart designed for moving folding chairs rolled over her foot. Her expert was a mechanical engineer with over thirty years' experience in the furniture industry, and he had worked on wheel placement on moveable furniture.¹⁴² He opined that the defendant could have made several changes to improve the safety of the cart in question.¹⁴³ He also opined

139. Proceeding on both design defect and warning defect claims can create something of an evidentiary conundrum, especially if the plaintiff suggests a reasonable alternative design. If one witness says there's a serious design defect for which a safer alternative exists, and another says "yes, yes, but the defendant should have warned against the problems caused by the defective design," the testimony is not completely consistent. *Bullock v. Volkswagen Group of America, Inc.*, 107 F. Supp. 3d 1305 (M.D. Ga. 2015), is an extreme example of the logical problems potentially created by asserting both a defect claim and a warning claim. The court held that even without a warnings expert the plaintiffs could assert a failure-to-warn claim because the defendant knew a defective turbo-charger might cause unexpected acceleration. *Id.* at 1316. Just what good such a warning might do went completely unexplained. The plaintiff did have two experts on the nature of the defect, and apparently neither of them provided an opinion about how a warning might ameliorate it. On the other hand, when a defect can, in fact, be remedied with a warning, the two kinds of defect claims don't conflict, and in some circumstances the warning claim can proceed on its own. *See, e.g.*, *Lara v. Delta Int'l Mach. Corp.*, 174 F. Supp. 3d 719 (E.D.N.Y. 2016) (in which the court excluded the plaintiff's design defect expert, but allowed the warning expert to testify).

140. *See In re FEMA Trailer Formaldehyde Prod. Liab. Litig.*, No. MDL 07-1873, 2009 WL 3247967, at *2 (E.D. La. Oct. 6, 2009) (court held there was no need for expert testimony on "the issue of the adequacy of a particular warning" because such a conclusion "involves a common sense assessment 'within the realm of the average juror's knowledge and experience'"); *McSwain v. Sunrise Med., Inc.*, Civ. Action No. 2:08cv136KS-MTP, 2010 WL 200004, at *8 (S.D. Miss. Jan. 14, 2010) (plaintiff proffered expert testimony that based on expert's experience, he knew "that a warning affixed to a product is more effective than a warning inside a manual"; court excluded the testimony because it was "well within the grasp of jurors who have undoubtedly experienced both types of warnings in their daily lives"). *Cf. Bradley v. Ameristep, Inc.*, 800 F.3d 205, 213 (6th Cir. 2015) (because consumer-expectations test applied in case about treestand ratchet straps, plaintiff "was entitled to have a jury employ its own sense of whether the relevant warnings provided the ordinary consumer with knowledge of how to effectively avoid or minimize the risks associated with the . . . straps").

141. 502 F. Supp. 2d 1025 (W.D. Mo. 2007).

142. *Id.* at 1027.

143. *Id.* at 1028-29.

that the “warning label was inadequate in color and wording [because it did not specify] the hazard, how to avoid the hazard and the consequences in not avoiding the hazard of the casters rolling over the feet during use.”¹⁴⁴ The court allowed him to testify on both defect and warnings, and on the latter held that he “was not talking ‘off the cuff’ but rather contrasting the existing ‘Notice’ on the cart with the ANSI standard and dozens of warnings he has previously drafted for other furniture himself.”¹⁴⁵

Still other courts allow testimony from engineers on the *need* for a warning, but not on the substance, appearance, or placement of the warning. In *Pineda v. Ford Motor Co.*,¹⁴⁶ for example, the plaintiff was an automotive repair technician who was injured while installing a replacement lift gate on a sport utility vehicle. His only expert, an engineer, testified that the manufacturer’s service manual should have laid out better step-by-step instructions for doing the installation, and that the “manual should have contained an explicit warning [about the need to follow those instructions].”¹⁴⁷ In reversing the trial court’s exclusion of this testimony, the Third Circuit held that while the engineer might not have been the best expert on warnings, his opinions should have been admitted, including testimony about a later warning issued by the defendant, as an example of what kind of language would be acceptable.¹⁴⁸ The court noted that the expert “did not purport to opine on how the warning should be worded or how it should appear in order to effectively convey its message to an automobile technician. He only testified that . . . a warning was necessary to alert a technician to the potential problem.”¹⁴⁹ Neither the engineering process nor the scientific method has much bearing on determining the admissibility of warning testimony in a case like *Pineda*.

144. *Id.* at 1029.

145. *Id.* at 1032. *See also* Tassin v. Sears, Roebuck & Co., 946 F. Supp. 1241, 1253 (M.D. La. 1996) (“The Court does not believe that a design engineer with experience in both product design and in preparing manuals and warnings must scientifically test an alternative warning before he can opine that a warning is defective.”); Weese v. Black & Decker, Inc., No. 05-3031-CV-S-RED, 2007 WL 4618526 at *1 (W.D. Mo. Sept. 11, 2007) (Expert allowed to testify on both design and warnings where he had “drafted instruction manuals ‘forty or fifty times, probably.’ He wrote warnings for products he designed, and he wrote warnings that are contained in service manuals.”).

146. 520 F.3d 237 (3d Cir. 2008).

147. *Id.* at 245.

148. *Id.* at 245–46.

149. *Id.* at 245. *See also* Miller v. BGHA, Inc., Civ. Action No. 19–129, 2021 WL 2681277, at *3 (E.D. Pa. June 30, 2021) (the court cited *Pineda* and held that an “expert need not ‘opine on how the warning should be worded or how it should appear’ to testify that the instructions might result in failure or that a warning is necessary”); Schaaf v. Caterpillar, Inc., 286 F. Supp. 2d 1070, 1074 (D.N.D. 2003) (expert allowed to testify about lack of warnings but not about “the appropriate size, shape, color, content or location of any such warning signs”); Furry v. Bielomatik, Inc., 32 Fed. App’x 882, 883 (9th Cir. 2002) (safety engineer could give both design and warning opinions at conceptual level, without details).

Causation

Proof of defect and proof of causation are often bound together in products liability cases. A plaintiff who proves (or fails to prove) one element almost automatically proves (or fails to prove) the other. In *Lapsley v. Xtek, Inc.*,¹⁵⁰ for example, the defendant did not dispute that a spurt of grease caused the plaintiff's injury. The question was whether a design defect caused the grease jet. The plaintiff's successful proof of the defect necessarily included proof of the cause.¹⁵¹ By way of contrast, in *Nease v. Ford Motor Co.*,¹⁵² the court rejected the plaintiffs' claim that a pickup truck crash occurred because a defectively designed accelerator cable had stuck and prevented the throttle from closing. Their expert hypothesized that dirt had entered the cable housing and caused the cable to bind, but he conceded it worked properly when he examined it, and that he'd conducted no tests to corroborate his hypothesis.¹⁵³ Failure to prove the cause necessarily entailed failure to prove the defect.¹⁵⁴

Although evidence of defect and cause are intertwined, the emphasis is on defect in most cases involving engineering experts, and the causation question is addressed secondarily or by implication. In cases about pharmaceutical products, medical devices, or exposure to allegedly toxic chemicals, the emphasis often reverses, and causation becomes the primary focus, but the experts typically are scientists or physicians rather than engineers. *Daubert*, which involved the drug Bendectin, is perhaps the most famous pharmaceutical example.

When a plaintiff attempts to use an engineering expert to establish medical causation in a medical products case, courts generally exclude the testimony. For example, in *Bayes v. Biomet, Inc.*,¹⁵⁵ the court admitted testimony from the plaintiffs' biomechanical engineering expert on what caused wear in a hip implant,¹⁵⁶

150. 689 F.3d 802 (7th Cir. 2012).

151. The reverse, however, was not true. Proof of the cause did not establish a defect because without showing there was a reasonable alternative design, the plaintiff would have lost. In fact, his expert explained that a design actually used by the defendant both before and after the accident would have prevented the spurting grease, or at least reduced its velocity (*Lapsley*, 689 F.3d at 807), which is why the plaintiff prevailed. *See also* *Williams v. Tristar Prods., Inc.*, 418 F. Supp. 3d 1212 (M.D. Ga. 2019) (pressure cooker lid popped off); *Bullock v. Volkswagen Grp. of Am., Inc.*, 107 F. Supp. 3d 1305 (M.D. Ga. 2015) (defective turbocharger caused uncontrollable acceleration). *Cf.* *Lara v. Delta Int'l Mach. Corp.*, 174 F. Supp. 3d 719 (E.D.N.Y. 2016) (injury clearly caused by saw; court rejected design defect expert but admitted testimony from failure-to-warn expert).

152. 848 F.3d 219 (4th Cir. 2017).

153. *Id.* at 231–32.

154. *See also* *Elkharoubi v. Six Flags Am. LP*, Civ. Action No. TJS-17-2169, 2020 WL 1043304 (D. Md. Mar. 4, 2020) (finding no evidence of defect in trap door on a water slide entailed finding no causation). *Cf.* *Colon ex rel. Molina v. BIC USA, Inc.*, 199 F. Supp. 2d 53 (S.D.N.Y. 2001) (although cigarette lighter clearly caused plaintiff's injuries, without evidence of a reasonable alternative design there was no product defect, and hence no cause of action).

155. Case No. 4:13-cv-00800-SRC, 2020 WL 5594059 (E.D. Mo. Sept. 18, 2020).

156. *Id.* at *5.

but her medical causation opinions were excluded because she was not a medical doctor.¹⁵⁷

Occasionally, the mix of engineering and medical experts occurs in the context of a non-medical product case. *Campbell v. Fawber*,¹⁵⁸ which involved injuries allegedly caused when the roof of an SUV crushed in an accident, provides an outstanding example of how engineers can address complex causation questions in the context of products liability litigation. The first of three engineering experts in *Campbell* had a degree in physics, but taught engineering at the National Crash Analysis Center at George Washington University.¹⁵⁹ He was allowed to testify about both the defective design of the SUV and how the structural failure of the roof caused the plaintiff's injuries.¹⁶⁰ The second expert had a Ph.D. in applied mechanics and had "taught courses on engineering physics and mechanics, and strength of materials, and limit analysis of structures."¹⁶¹ He offered an opinion on the strength of the SUV's roof, and the vehicle's propensity to roll over.¹⁶² The third engineering expert was a biomedical engineer who had "taught courses in the areas of applied mechanics, materials science, biomechanics, biomaterials, ergonomics, occupational health and safety, strength of materials, and orthopaedic engineering."¹⁶³ He testified that the manufacturer could have designed the roof so the amount of collapse would have been greatly reduced, and that a poorly designed seat belt may also have contributed to the plaintiff's injuries.¹⁶⁴

In explaining its decision to admit the testimony of all three engineers, the *Campbell* court went to some length to distinguish their work from that done by experts in other cases who were excluded. The first engineering expert "uses established systems analysis techniques to form his opinion, evaluating 'crash statistics, vehicle dynamics, occupant kinematics, injury susceptibility and mechanism, crashworthiness, and occupant protection.' He examines a number of materials, including those published in peer-reviewed journals. Most importantly, however, [he] relies on [information about other vehicles with a roof structure similar to the SUV, and with alternative designs]."¹⁶⁵ The second expert reviewed extensive information and concluded that "the roof intrusion over [the plaintiff's] seat would have subjected her to as much as twice [the] dynamic deflection as her body could withstand."¹⁶⁶ Similarly, the third expert "observed the crash location,

157. *Id.* at *6.

158. 975 F. Supp. 2d 485 (M.D. Pa. 2013).

159. *Id.* at 491.

160. *Id.* at 492.

161. *Id.* at 493.

162. *Id.*

163. *Id.* at 494.

164. *Id.*

165. *Id.* at 498.

166. *Id.* at 499.

and took measurements in order to reconstruct the accident scene [and explained that the SUV had a higher-than-acceptable propensity to roll over].”¹⁶⁷

The defendant argued that no expert had conducted a test of a modified SUV to show that remedying the supposed defects would have prevented the plaintiff’s injuries. The court rejected this rote invocation of testing. To require such tests would make plaintiffs “prove their case twice.”¹⁶⁸ The court also noted that the factors listed in *Daubert* do not constitute a definitive checklist or test for admissibility.¹⁶⁹ Thus, *Campbell* is another example of how engineering design evaluation rather than scientific testing of hypotheses can be the basis for admissible causation testimony.

There are cases, however, in which causation testimony from an engineering expert is properly analyzed based on hypothesis testing criteria. Many fire investigation cases fall into this category. *Werth v. Hill-Rom, Inc.*¹⁷⁰ is an excellent example that also involved products liability. The plaintiff was a newborn infant who required oxygen therapy. He was placed under an oxygen hood in a bassinet with a baby warmer over him.¹⁷¹ When a fire erupted in the bassinet, the hospital and his parents blamed it on “a hot particle [that] had broken from the baby warmer and fallen into the bassinet, igniting the flammable materials (blankets and the like) below,”¹⁷² but the court excluded expert testimony on this theory of causation. Citing a publication from the National Fire Protection Association, it found that “once an investigator has developed a hypothesis about a fire’s cause, [a key step] is *testing* that hypothesis. Yet, nothing in the Report or the Supplemental Report indicates that the team tested its hypothesis at all. Indeed, as [one investigator] candidly acknowledged in his deposition, it never did.”¹⁷³

Duty of Care

In products liability cases based on the legal theory of negligence rather than strict liability or warranty, a plaintiff has to establish that the defendant breached a duty of care, and engineers sometimes testify about the duty of care as well as product defect or causation. In *Crawford v. ITW Food Equipment Group, LLC*,¹⁷⁴ the plaintiff lost an arm “when it came into contact with the unguarded blade of one of [the defendant’s] commercial meat saws.”¹⁷⁵ His engineering expert testified about

167. *Id.*

168. *Id.*

169. *Id.*

170. 856 F. Supp. 2d 1051 (D. Minn. 2012).

171. *Id.* at 1053.

172. *Id.*

173. *Id.* at 1060.

174. 977 F.3d 1331 (11th Cir. 2020).

175. *Id.* at 1336.

alternative designs of which the defendant should have known, and this testimony sufficed to establish negligent design.¹⁷⁶ Obviously, testimony about what the defendant should have known is a matter of engineering knowledge, and *Daubert*'s admissibility criteria for scientific evidence would not apply.

Non-Products Liability Cases

Contract and Construction Litigation Cases

Engineers are often called as experts when disputes arise about contract compliance, especially in the context of construction contracts.¹⁷⁷ For example, *Waste Management of Louisiana, L.L.C. v. Jefferson Parish*¹⁷⁸ involved a complex dispute between Jefferson Parish, Louisiana, and Waste Management, the company that operated, managed, and maintained the parish's sanitary landfill.¹⁷⁹ The core issues in the case related to (1) what work the plaintiff, Waste Management, and its subcontractor and counterclaim plaintiff, T & K Construction, were supposed to do, (2) when it was to be done, and (3) how the parish might have interfered with that work. The parish retained a civil engineer to serve as an expert on these issues, and the two plaintiffs moved to exclude his testimony.

The plaintiffs did not challenge the expert's methodology or theories, or the data he considered, "except to the extent that they challenge his interpretations of provisions of the [contract with Waste Management]."¹⁸⁰ Though the court recited the *Daubert* factors, it made no real attempt to apply them, and ultimately held that the expert had "explained his methodology, which involved a careful review of all of the relevant pleadings and documents provided by the parties, as well as an interview with the Landfill engineer and a visit to the Landfill. [His] expert report reveals that he has more than 30 years of engineering experience across the fields of sewerage systems, site engineering, land surveying and structural engineering, among others, albeit not seemingly focusing on landfills. [Though he] may lack expertise on certain questions of landfill construction, . . . his overall experience suggests that Waste Management and T & K have failed to set forth that any such failing disqualifies [his] expert report in its entirety."¹⁸¹

The court was concerned, however, that the expert had strayed from engineering testimony into contract interpretation. "The Court agrees that [the]

176. *Id.* at 1343–44.

177. For a good discussion of construction industry *Daubert* cases, see C.J. Heffernan et al., *Defending and Asserting Daubert Challenges in Construction Disputes*, 32 *The Construction Lawyer* (Spring 2012).

178. Civ. Action No. 13–6764, 2015 WL 5798029 (E.D. La. Oct. 5, 2015).

179. *Id.* at *1.

180. *Id.* at *12.

181. *Id.*

expert report is rife with legal conclusions and contractual interpretations. . . . However, the Court finds that although many of the opinions offered in [the] expert report are inadmissible, [the expert] . . . may testify that “[a]s a national landfill operator, Waste Management should have been aware of the problems that gas could cause in capping work and should have made provisions for it in the plans they developed.”¹⁸²

Waste Management is a prime example of how the *Daubert* factors do not properly apply to some engineering testimony, and other contract cases further illustrate this point. In *First Assembly of God Church v. Fondren*,¹⁸³ the issues were whether a construction company had complied with drawings and specifications, and whether failure to comply had resulted in the collapse of a church’s roof. The defendant raised about every argument possible in unsuccessfully challenging the plaintiff’s engineering expert.¹⁸⁴ On the facts of the case, however, the broadside attack on the expert verged on the frivolous. The expert was “a registered public engineer in nine states . . . and [had] actively practiced engineering for [about 29 years].”¹⁸⁵ He had “previously conducted investigations into numerous other building design and construction defects as well as building collapses,” co-authored articles, and “testified in numerous cases.”¹⁸⁶ Here, the expert opined that “[b]ased on my knowledge of the behavior of rigid frame metal buildings, I determined that braces should have been installed at the knee joints and that braces would have precluded the observed failure.”¹⁸⁷ The defendant did not raise “any specific objections to the methodology used by [the expert] . . . [and made] no argument as to why additional calculations on the bending strength of the steel frame [were] necessary.”¹⁸⁸

182. *Id.* at *13.

183. No. 5:02CV47, 2003 WL 25685226 (E.D. Tex. Jan. 8, 2003).

184. *Id.* at *3. There were eight supposed grounds for the motion to exclude: “(1) he is not qualified to give the opinions; (2) his theories have not and cannot be tested; (3) his testimony is based upon subjective interpretation rather than valid theory; (4) his theories have not been subjected to proper peer review and/or publication; (5) his opinions have a potential high rate of error; (6) he is attempting to testify to matters which are not scientifically correct and with no scientific basis for his opinions; (7) his opinions and conclusions should be excluded because they are incorrect, and the analysis used to reach the opinions and conclusions are not reliable; and (8) he is not a licensed engineer in Texas.”

185. *Id.* at *5.

186. *Id.*

187. *Id.* at *6. *Cf. Allegra v. Luxottica Retail N. Am.*, 341 F.R.D. 373, 438–42 (E.D.N.Y. 2022) (plaintiffs claimed defendant manufacturer of eyeglasses had made deceptive misrepresentations about the significance of the precision with which its lenses were made; court held an engineer could testify for plaintiffs based on his review of literature about the lens-making machines at issue).

188. *Id.* See also *United States ex rel. Poong Lim/Pert Joint Venture v. Dick Pac./Ghemm Joint Venture*, 2006 WL 5230015, at *3 (D. Alaska Mar. 2, 2006) (in a dispute between a contractor and subcontractor, the court rejected the plaintiff’s motion to exclude the defense experts, holding that in “cases where the subject of the expert testimony is non-scientific, as the case at bar, education,

Waste Management and *First Assembly of God* do not establish a blanket rule for experts in construction cases. Generally, it does take more than educational background and experience for engineering testimony to be admissible. The exclusion of two architecture experts in *Cross Creek Multifamily, LLC v. ICI Construction, Inc.*¹⁸⁹ illustrates all too well that there has to be some rational basis for an expert's opinion. The issue before the court was how to apportion responsibility for construction defect damages among a project's architect, general contractor, and several subcontractors. When asked at his deposition how he had come to assign 64% responsibility to the architect and engineers, the expert for one defendant essentially answered that he had used his "training, education, and experience, [and] look[ed] at all the documentation. . . . That's how I saw it, and that's how I distributed it."¹⁹⁰ Upon further questioning, he admitted there was no "exact formula by [the American Institute of Architects] or anyone that works in these situations."¹⁹¹ He also testified that he'd never done such an apportionment before, that it was "a very unusual and difficult situation," and that there existed no "specific book or reference or industry material . . . that tells an architect how to do [it]."¹⁹² The expert for another defendant fared no better. He admitted that he did not have all the relevant information when he initially opined, and when provided additional information during his deposition, he said he *probably* would change his apportionment.¹⁹³

Intellectual Property Cases

Engineers in copyright cases

Engineering testimony does not figure in most copyright cases, but there are exceptions. In *Vargas v. Transeau*,¹⁹⁴ the plaintiffs claimed the defendants had copied from their sound recording and then digitally manipulated the sound to create a new and infringing recording.¹⁹⁵ To make their case, however, they had to explain why the two recordings were not identical. Thus, they sought to introduce testimony from "a purported sound engineer and sampling expert . . . who asserts that [one of the defendants] created [his recording] by importing drum

training and experience, not necessarily methodology or theory, are more appropriate benchmarks for determining reliability"); *Ferguson v. Erie Ins. Prop. & Cas. Co.*, Civ. Action No. 3:19-0810, 2021 WL 4392040, at *3 (S.D. W. Va. Sept. 24, 2021) (in dispute about what caused walls to crack, expert allowed to testify "as to his own personal observations and experience as an engineer").

189. Civ. Action No. 2:18-cv-83-KS-MTP, 2020 WL 8254812 (S.D. Miss. Dec. 1, 2020).

190. *Id.* at *2.

191. *Id.*

192. *Id.* at *2–*3.

193. *Id.* at *3.

194. 514 F. Supp. 2d 439 (S.D.N.Y. 2007).

195. *Id.* at 441.

beats from [the allegedly infringed recording].”¹⁹⁶ The court granted summary judgment after finding that the expert had failed “to substantiate his opinion. . . . He offers no explanation for what the tools [for sound manipulation] are; provides no significant detail regarding how they alter sound; and provides no demonstration of how a sound from [plaintiff’s recording] would be manipulated to achieve one of the sounds in [the other recording].”¹⁹⁷ The engineering issue in cases like *Vargas* would be how engineers develop and use tests to determine similarity of recordings.¹⁹⁸

Engineers in trademark cases

Engineers sometimes appear in trademark cases, though the experts more typically are people who conduct surveys to see if one party’s use of a logo or the like has caused confusion with another party’s logo or the like. In *Malletier v. Dooney & Bourke, Inc.*,¹⁹⁹ one issue was whether the defendant’s use of certain colors caused confusion about multicolor handbag logos.²⁰⁰ The court denied a motion to exclude an expert on colorimetry who was “a principal engineer in color technology, and a professor at Boston University, researching and teaching issues pertinent to vision and color.”²⁰¹ Again, the engineering issue would be the development and use of tests.

Engineers in patent cases

Most engineering expert testimony on intellectual property issues occurs in patent litigation. In the United States, patents are issued by the U.S. Patent and Trademark Office (USPTO) after an extensive application and examination process.²⁰² There are three kinds of patent. A utility patent “may be granted to anyone who invents or discovers a new and useful process, machine, article of manufacture, or composition of matter, or any new and useful improvements of

196. *Id.* at 445.

197. *Id.*

198. See also *SA Music LLC v. Apple, Inc.*, 592 F. Supp. 3d 869, 900–02 (N.D. Cal. 2022) (excluding testimony from expert software engineer who specialized in audio and music technology).

199. 525 F. Supp. 2d 558 (S.D.N.Y. 2007).

200. *Id.* at 570.

201. *Id.* at 641.

202. See Peter S. Menell et al., *Patent Case Management Judicial Guide*, Third Edition (Federal Judicial Center 2016), pp. 14–9 to 14–29.

these.”²⁰³ A utility patent may cover a business method.²⁰⁴ A design patent may be granted to “anyone who has invented a new, original ornamental design for an article of manufacture.”²⁰⁵ A plant patent may be granted to “anyone inventing or discovering and asexually reproducing any distinct and new variety of plant.”²⁰⁶ Utility patents far and away constitute the majority of patents granted,²⁰⁷ and they are the primary focus here.

Every patent includes a specification that describes the patented invention and concludes with what are called claims. Claims “are commonly analogized to the ‘metes and bounds’ of a property deed and serve the same purpose: to delineate the scope of the asset which, in the patent context, is an invention.”²⁰⁸ Importantly, a patent does not give the holder the right to make or sell anything. Rather, each “claim represents the legal right to *exclude others* from making, using, selling, offering to sell, importing, or offering to import the claimed process, machine, manufacture, or composition of matter.”²⁰⁹ “A patent may, and often does, contain many claims, which usually become increasingly specific” as they progress from first claim to last.²¹⁰

A simple hypothetical illustrates how claims work and what rights they confer if the USPTO grants a patent. Suppose Inventor A has a patent with two claims: (1) a drink container comprising a cylinder closed at one end and open at the other, and (2) a drink container as in Claim One wherein the closed end cylinder is made of ceramic material. Claim One, which would be considered an independent claim, would cover, or “read on” almost any drinking glass, except, for example, a non-cylindrical glass that’s wider at the top than at the bottom. Claim Two, which would be called a dependent claim, would only read on a drinking “glass” made of ceramic material.²¹¹

Now suppose Inventor B comes up with a new idea that would make it easier to use one of the cylindrical drink containers for hot beverages, like coffee.

203. See United States Patent and Trademark Office, “Applying for Patents,” available at <https://perma.cc/JY3T-Z83H> (accessed Mar. 6, 2024).

204. *State Street Bank & Trust Co. v. Signature Fin. Grp., Inc.*, 149 F.3d 1368 (Fed. Cir. 1998).

205. See USPTO, “Applying for Patents,” *supra* note 203.

206. *Id.*

207. Menell, Patent Case Management Judicial Guide, at 1–13.

208. *Id.* at 14–7 to 14–8.

209. *Id.* at 14–8 (emphasis added) & 35 U.S.C. § 271(a).

210. Menell, Patent Case Management Judicial Guide, at 14–8.

211. Note that in patent law, the term “comprising” simply means having *at least* the elements in the claim. See *CIAS, Inc. v. Alliance Gaming Corp.*, 504 F.3d 1356, 1360 (Fed. Cir. 2007) (“In the patent claim context the term ‘comprising’ is well understood to mean ‘including but not limited to.’”), cited in *In re Skvorecz*, 580 F.3d 1262, 1267 (Fed. Cir. 2009). *Cf.* *Guardant Health, Inc. v. Foundation Medicine, Inc.*, Civ. Action No. 17-1623-LPS-CJB, 2019 WL 5677748 at *15 (D. Del. Nov. 1, 2019) (making clear that an independent claim must be construed to include a dependent claim).

Inventor B obtains a patent with a single claim: a drink container comprising a ceramic cylinder closed at one end and open at the other and a ceramic handle attached to said cylinder. The claim in this patent would read on a coffee mug, at least one cylindrical in shape. Can Inventor B now start making and selling cylindrical coffee mugs with handles? No, because doing so would infringe on the claims in Inventor A's patent. Inventor A, however, could not make or sell the mugs either, because doing so would infringe on Inventor B's patent. Would it then be impossible to make and sell coffee mugs in our hypothetical world? As a practical matter, probably not, because the two patent holders likely would enter into a cross-licensing agreement that would let them both make mugs.²¹²

Patent litigation almost always arises when a patent holder believes a competitor is making or selling something that infringes on the patent, or when the competitor seeks a declaratory judgment that its "accused" process, product, or method does not infringe. Infringement means at least one claim is infringed;²¹³ that is, at least one claim reads on the allegedly infringing process, product, or method. Once litigation has commenced, there are three principal issues on which experts may testify: (1) claim construction, (2) infringement or non-infringement, and (3) validity or invalidity.

Claim construction Claim construction, or "the process by which courts interpret patent claims, represents one of the most distinctive aspects of patent litigation."²¹⁴ "When patentees seek to enforce their rights in court, the interpretation of patent claim boundaries guides both infringement and validity analysis."²¹⁵ In the coffee mug hypothetical, for example, the manufacturer of an accused cup that tapered from a wide opening at the top to a smaller diameter at the bottom might seek construction of the claim in Inventor B's patent as covering only cylindrical cups. In such a case, the claim construction likely would be dispositive, and under the U.S. Supreme Court's 1996 decision in *Markman v. Westview Instruments*,²¹⁶ the judge, rather than the jury, does the analysis of any claims at issue.

Justice Breyer's opinion in *Markman* recognized that because a trial court has to consider at least some evidence in addition to the language of the claims, interpretation of a claim is not the kind of purely legal question typically addressed by a trial judge rather than the jury. Construing "a term of art following receipt of

212. For a real-life example of cross-licensing, see *Intel Corp. v ULSI System Technology, Inc.*, 995 F.2d 1566 (Fed. Cir. 1993).

213. *Grober v. Mako Prods., Inc.*, 686 F.3d 1335, 1344 (Fed. Cir. 2012).

214. Menell, Patent Case Management Judicial Guide, *supra* note 202, at 5–3.

215. J. Jonas Anderson & Peter S. Menell, *Informal Deference: A Historical, Empirical, and Normative Analysis of Patent Claim Construction*, 108 Northwestern Univ. L. Rev. 1, 3–4 (2014).

216. 517 U.S. 370, 384–88 (1996).

evidence” was termed a “mongrel practice.”²¹⁷ Nonetheless, taking into account precedents in prior cases and “the relative interpretive skills of judges and juries and the statutory policies that ought to be furthered by the allocation [between judge and jury],”²¹⁸ the Court held “there is sufficient reason to treat construction of terms of art like many other responsibilities that we cede to a judge in the normal course of trial, notwithstanding its evidentiary underpinnings.”²¹⁹

Only a few months after *Markman*, the Federal Circuit, in *Vitronics Corp. v. Conceptronic, Inc.*,²²⁰ held that a trial judge had erred in relying on expert testimony in interpreting a claim. The Federal Circuit distinguished between intrinsic evidence (the patent file and history) and extrinsic evidence (e.g., expert testimony, inventor testimony, dictionaries, and technical treatises and articles).²²¹ “[E]xtrinsic evidence in general, and expert testimony in particular, may be used only to help the court come to the proper understanding of the claims; it may not be used to vary or contradict the claim language.”²²² Given this rule, expert testimony during the claim construction stage of a litigated case is less frequent than during the trial, when experts often testify about infringement and validity.²²³

Moreover, when experts do testify on claim construction, their opinions typically derive from their experience, and not from any engineering analysis or scientific hypothesis testing. Thus, challenges to the admissibility of their testimony focus mostly on their education, training, and experience. *Pentair Water Pool and Spa, Inc. v. Hayward Industries, Inc.*²²⁴ is a good example of how courts evaluate such testimony. The case involved heaters for swimming pools and spas. During the *Markman* hearing, the defendant tried to introduce testimony from a materials engineering professor. The court held the professor’s testimony largely inadmissible because the expert was not someone with adequate “skill in the

217. *Id.* at 378.

218. *Id.* at 384.

219. *Id.* at 390.

220. 90 F.3d 1576 (Fed. Cir. 1996).

221. *Id.* at 1584.

222. *Id.*

223. The allocation of claim construction to the judge quickly led to the practice of holding pretrial “*Markman* hearings.” Such hearings promote settlement and provide a basis for summary judgment. J. Jonas Anderson & Peter S. Menell, *Informal Deference: A Historical, Empirical, and Normative Analysis of Patent Claim Construction*, 108 Northwestern Univ. L. Rev. 25 (2014). These hearings also facilitate the development of expert reports on the issues of infringement or validity, because they clarify how the claims in a patent have to be interpreted. *Id.* Because *Markman* hearings occur well before trial, and sometimes even before discovery, they often involve no expert testimony at all. J. Michael Jakes, *Using an Expert at a Markman Hearing: Practical and Tactical Considerations*, IP Litigator, Aug. 2002 at 2 (“Once the primary means for presenting claim-construction issues to juries, expert testimony may or may not be allowed by the courts today.”), available at <https://perma.cc/GLH3-UXVG>.

224. Case No. CV 11-10280-GW(FMOx), 2012 WL 12887686 (C.D. Cal. Dec. 12, 2012).

[relevant] art.”²²⁵ The court found that while he was “clearly a distinguished Professor of Engineering, [he was not] a person of ordinary skill in the art in the field to which the . . . patent pertains. He admitted as much at his deposition where he said ‘I’m not of ordinary skill, but I know who such people are. I teach them and have taught them for years.’”²²⁶

Infringement Unlike during claim construction, expert testimony is allowed once a patent case moves on to a determination of whether infringement (when at least one claim reads on the defendant’s accused product) has occurred. The admissibility review, however, is very different from cases in other areas, where the focus is typically on methodology and testing and/or analysis based on accepted principles and standards. As an initial matter, expert testimony that contradicts a court’s construction of the patent will not be admitted. As the court held in *Exergen Corp. v. Wal-Mart Stores, Inc.*,²²⁷ once “a district court has construed the relevant claim terms, and unless altered by the district court, then that legal determination governs for purposes of trial.”²²⁸

When the actual merits of expert testimony on infringement are addressed, the scrutiny is often less strict than in products liability cases. In *Fontem Ventures, B.V. v. NJOY, Inc.*,²²⁹ a case involving several patents for e-cigarette devices, the plaintiff’s expert on infringement was a mechanical engineer with “40 years’ experience in the design, development, and manufacture of electromechanical devices, both as an employee and as a consultant.”²³⁰ For each allegedly infringing device he “provide[d] photographs of the accused products annotated with numerals in the ‘Infringement Analysis’ and brackets or arrows pointing to portions of the accused products, referencing numerals from [various exhibits].”²³¹ His conclusion was essentially “look at the patent and look at what the defendants have done.” Neither the engineering design process nor the scientific method was involved.

225. *Id.* at *11.

226. *Id.* See also *Daiichi Sankyo Co. v. Apotex, Inc.*, 501 F.3d 1254, 1256 (Fed. Cir. 2007) (listing factors that a court can consider for determining if an expert has the requisite level of ordinary skill in the art at issue).

227. 575 F.3d 1312 (Fed. Cir. 2009).

228. *Id.* at 1321.

229. CV 14-1645-GW(MRWx), 2015 WL 12743861 (C.D. Cal. Oct. 22, 2015).

230. *Id.* at *3.

231. *Id.*

In ruling the expert's testimony admissible, the court held that

a patent infringement expert's opinions as to certain uncomplicated elements can be based on a visual inspection and operation of the accused products, as well as the expert's knowledge, skill, experience, and education. The traditional and appropriate means of attacking 'shaky' or simplistic but admissible evidence is by vigorous cross-examination, presentation of contrary evidence, and careful instruction on the burden of proof.²³²

The non-infringement testimony of the defendant's expert was excluded, however, because at his deposition he

repeatedly pointed out that he did not know various details about the accused . . . products, [and explained] that his lack of knowledge stemmed from his opinion that [the plaintiff's expert] report failed to inform him about such facts. While he may rely on the alleged weakness of [the opposing expert's] report in his rebuttal testimony, he must actually have a sufficient grasp of the relevant facts to form non-infringement opinions of his own.²³³

Validity A defendant that has arguably infringed claims in a patent can still escape liability by proving that the relevant claims were invalid. That is, the patent should not have been granted with those claims in the first place. Establishing invalidity is not easy, however. Under 35 U.S.C. § 282, a "patent shall be presumed valid. Each claim of a patent . . . shall be presumed valid independently of the validity of other claims . . ." Moreover, the statute provides that the "burden of establishing invalidity of a patent or any claim thereof shall rest on

232. *Fontem Ventures*, 2015 WL 12743861, at *7. See also *Gatearm Techs., Inc. v. Access Masters, LLC*, Case No. 14-62697-CIV-SCOLA/OTAZO-REYES, 2020 WL 6808670, at *14 (S.D. Fla. Apr. 30, 2020) (defendant's expert testimony on non-infringement admitted based on his comparison of defendant's original infringing product with a new design, which he concluded was very different).

233. *Fontem Ventures*, 2015 WL 12743861, at *8. See also *DataQuill Ltd. v. Handspring, Inc.*, No. 01 C 4635, 2003 WL 737785, at *2-4 (N.D. Ill. Feb. 28, 2003) (in a case involving data entry devices, the court held that "operating and examining the accused devices (assuming it is done against the backdrop of a proper claim construction) is a reliable method of conducting an infringement analysis," but nonetheless excluded expert's testimony because he had not properly construed the claims). Cf. *Fuma Int'l LLC v. R.J. Reynolds Vapor Co.*, 1:19-CV-260, 1:19-CV-660, 2021 WL 4820738, at *2 (M.D.N.C. Oct. 15, 2021) (another case involving e-cigarette devices). In *Fuma*, the court held the plaintiff's expert testimony admissible to the extent it was "based on technical analysis on several different points. He discusses technical design similarities between Fuma's patented e-cigarette design and RJR's Solo and Ciro products, such as a transverse heating element and central airflow passageway. . . . He also uses a technical analysis to explain his opinion that RJR's argument that Fuma's e-cigarette design was different than its own is not valid, and that RJR's portrayal of Fuma's e-cigarette design could not electrically function." But see *Elder v. Tanner*, 205 F.R.D. 190 (E.D. Tex. 2001), in which the court held it "is not sufficient simply to list the resources [experts] utilized and then state an ultimate opinion without some discussion of their thought process." *Id.* at 194.

the party asserting such invalidity,”²³⁴ which requires clear and convincing evidence.²³⁵

There are a number of grounds for invalidity, but a 2012 scientific review paper prepared by the Morgan, Lewis & Bockius law firm on invalidation of patents by the Federal Circuit indicates that anticipation and obviousness are far and away the most common reasons.²³⁶ Anticipation is addressed in 35 U.S.C. § 102,²³⁷ and obviousness in 35 U.S.C. § 103.²³⁸ Anticipation essentially means a single prior patent or publication covered the claims at issue, or the invention was available to the public before the filing date for the patent.²³⁹ Obviousness essentially means that a person skilled in the relevant art could easily have come up with the invention based on available information.²⁴⁰ Note that a claim can be non-obvious yet still anticipated.²⁴¹

When expert testimony is introduced regarding validity, it usually involves only a comparison of the claims in the patent at issue with relevant “prior art.” Thus, admissibility decisions typically come down to the expert’s qualifications.

234. *Id.*

235. *Schumer v. Lab. Computer Sys., Inc.*, 308 F.3d 1304, 1315 (Fed. Cir. 2002) (“Evidence of invalidity must be clear as well as convincing.”).

236. White Paper Report: United States Patent Invalidation Study 2012, available at https://www.morganlewis.com/~media/files/publication/presentation/speech/smyth_uspatentininvalidity_sept12 (accessed July 14, 2024).

237. Section 102, 35 U.S.C., provides that a person is not entitled to a patent if “(1) the claimed invention was patented, described in a printed publication, or in public use, on sale, or otherwise available to the public before the effective filing date of the claimed invention; or (2) the claimed invention was described in a patent issued under section 151, or in an application for patent published or deemed published under section 122(b), in which the patent or application, as the case may be, names another inventor and was effectively filed before the effective filing date of the claimed invention.”

238. Section 103, 35 U.S.C., provides that a “patent for a claimed invention may not be obtained, notwithstanding that the claimed invention is not identically disclosed as set forth in section 102, if the differences between the claimed invention and the prior art are such that the claimed invention as a whole would have been obvious before the effective filing date of the claimed invention to a person having ordinary skill in the art to which the claimed invention pertains.”

239. *See Therasense, Inc. v. Becton, Dickinson*, Case Nos. C04-02123 MJJ, C04-03327 MJJ, C04-03732 MJJ, 2007 WL 2028197 at *3 (N.D. Cal. July 10, 2007) (“A claim is anticipated under 35 U.S.C. § 102 only if each and every limitation of the claim is disclosed in a single prior art reference.”).

240. *See Izzo Golf, Inc. v. King Par Golf Inc.*, 561 F. Supp. 2d 334, 338 (W.D.N.Y. 2008) (“A claim is **obvious** if the patented invention would have been **obvious** to a person having ordinary skill in the art to which the subject matter of the patent pertains, at the time the invention was disclosed.”) (emphasis in original).

241. *Cohesive Techs., Inc. v. Waters Corp.*, 543 F.3d 1351, 1364 (Fed. Cir. 2008) (“While it is commonly understood that prior art references that anticipate a claim will usually render that claim obvious, it is not necessarily true that a verdict of nonobviousness forecloses anticipation. The tests for anticipation and obviousness are different. . . . Obviousness can be proven by combining existing prior art references, while anticipation requires all elements of a claim to be disclosed within a single reference. . . . ‘[I]t does not follow that every technically anticipated invention would also have been obvious.’”) citing *In re Fracalossi*, 681 F.2d 792, 796 (CCPA 1982) (Miller, J., concurring).

Mobile Medical International Corp. v. Advanced Mobile Hospital Systems, Inc. illustrates how courts analyze the admissibility of engineering testimony on obviousness.²⁴² The plaintiff company in the case brought a declaratory judgment action to invalidate the defendant's patent for a mobile medical unit. The plaintiff company's expert was a consultant in the field of specialty vehicle engineering, including medical vehicles. The defendant challenged the expert's experience and also argued that his opinion on obviousness was based on "hindsight bias." The court rejected a *Daubert* motion to exclude the expert, holding on the hindsight issue that his testimony about not using post-patent information sufficed.²⁴³ There was nothing about the application of the *Daubert* factors for scientific testimony or about the engineering design process.

*Meridian Manufacturing, Inc. v. C&B Manufacturing, Inc.*²⁴⁴ provides an example of a more detailed analysis of expert testimony on obviousness, but while the court initially discussed *Daubert* at some length,²⁴⁵ the ultimate decision on two challenged engineering experts was based mostly on the engineering analyses they had done. The invention at issue was an agricultural trailer specially designed for the transport of large bags or boxes of seeds to planters in the field.²⁴⁶ The defendant (alleged infringer) claimed the patent was "clearly an obvious combination of well-known elements from other patents."²⁴⁷ The court undertook a point-by-point discussion of each expert, and focused largely on their explanations for how they reached their conclusions. For example, the court found that at the beginning of the report prepared by the defendant's expert, he "discussed generally the geometric considerations and principles of guiding objects onto surfaces that go towards the 'common sense' category of motive to combine. While not determinative, and certainly subject to cross-examination and rebuttal, this [was] sufficient to permit [his] obviousness opinion to be admitted into evidence."²⁴⁸

Engineers in trade secret cases

Engineers also sometimes appear as expert witnesses in trade secret cases. Although the primary fact questions in such cases are usually about who stole or disclosed secret information, defendants sometimes will claim the information wasn't really a secret because anyone could easily have figured it out. In *Raytheon Co. v. Indigo Systems Corp.*,²⁴⁹ the defendant's expert opined that some supposed trade secrets

242. No. 2:07-cv-231, 2015 WL 778553 (D. Vt. Feb. 24, 2015).

243. *Id.* at *4.

244. 340 F. Supp. 3d 808 (N.D. Iowa 2018).

245. *Id.* at 829–30.

246. *Id.* at 823.

247. *Id.* at 828.

248. *Id.* at 834.

249. 598 F. Supp. 2d 817 (E.D. Tex. 2009).

could have been reverse engineered.²⁵⁰ He also suggested that other alleged trade secrets did not “meaningfully differ from information that is widely known in the industry.”²⁵¹ The court admitted his testimony with no detailed analysis.²⁵²

Environmental/Land Use Cases

The use of experts in cases about environmental or land use claims goes back almost 250 years, to *Folkes v. Chadd*,²⁵³ the English case generally considered the first example of allowing experts to provide opinion testimony. Today, engineers often testify in such cases. For example, in *Adler v. Elk Glenn, LLC*, the court admitted testimony from a geotechnical engineer about the cause of land subsidence under the plaintiff’s property.²⁵⁴ This expert had

consulted mine maps and websites, United States Geological Survey maps, topographic maps, excavated material, and rock strata visible in nearby areas. He also relied on his extensive experience with mine spoil to figure out the likely distribution of materials within the fill beneath the house. . . . He developed this opinion based on his extensive personal experience with mine spoil and his review of scholarly literature finding that mine spoil settlement worsens over time [and he visited the property].²⁵⁵

The fact that he had not performed readily available tests did not make his testimony unreliable because he had “followed the prevailing methods of his profession.”²⁵⁶ *Adler* is another example of how engineering analysis rather than scientific hypothesis testing can suffice to establish causation.

250. *Id.* at 821.

251. *Id.* at 821–22.

252. *Cf. Coda Development s.r.o. v. Goodyear Tire & Rubber Co.*, 565 F. Supp. 3d 979 (N.D. Ohio 2021), in which the court denied a motion to exclude the plaintiff’s expert on misappropriation of trade secret information about self-inflating tires because the defendant’s motion had focused on the expert’s knowledge and experience, and had “not presented a rigorous application of the teachings of *Daubert*.” *Id.* at 998. It is unclear how the scientific method might have been relevant.

253. 99 Eng. Rep. 589 (1782).

254. 986 F. Supp. 2d 851 (E.D. Ky. 2013).

255. *Id.* at 855.

256. *Id.* at 856. *See also* *Lake Charles Harbor & Terminal Dist. v. Reynolds Metal Co.*, Case No. 2:17-CV-01114, 2022 WL 838394 at *3 (W.D. La. Mar. 21, 2022) (expert on cleaning up landfills allowed to testify “based on his experience, expertise, and education”).

Failure Analysis Cases

The 1981 collapse of a hotel skywalk in Kansas City, discussed above, provides a classic example of both design and failure analysis. Science and hypothesis testing were not at issue.

*Georgia Power Co. v. Sure Flow Equipment, Inc.*²⁵⁷ is a more recent example of a failure analysis case. The plaintiff power company bought strainers from the defendant for use in removing debris from turbine steam cooling systems.²⁵⁸ They “failed in service, broke into pieces, went upstream into two turbine units and caused damage.”²⁵⁹ The central issue was the adequacy of the design.²⁶⁰ But while the court cited *Kumho* once, its analysis was based more on *Daubert* and a number of cases cautioning that the *Daubert* criteria are not absolute requirements. In the end, it admitted the disputed testimony from the plaintiff’s expert in “engineering, metallurgy, design of materials, and failure analysis”²⁶¹ after finding that it was “sufficiently reliable. His report, affidavits, and deposition testimony demonstrate that he has tested his theories, his methodology is supported by peer-reviewed research, and his conclusions are based on sufficient underlying facts and data.”²⁶² While this ruling seems well reasoned and correct, the emphasis on *Daubert* reflects the conflation of engineering and science sometimes seen in other cases.

Testing Cases

In some cases (most typically involving contract compliance), engineers simply testify about test results, where a measurement or some other parameter is at issue. In *Illinois National Insurance Co. v. Ace Stamping and Machine Co., Inc.*, for example, a central question was whether certain washers met hardness and flatness requirements.²⁶³ Testimony from the plaintiff’s expert about testing for these two characteristics was admitted. In *Cedar Petrochemicals, Inc. v. Dongbu Hannong Chemical Co., Ltd.*, the issue was whether phenol sold by the defendant to the plaintiff conformed to contract specifications.²⁶⁴ In particular, the plaintiff claimed the phenol was too dark, and relied on test results that measured the color in

257. Civ. Action No. 1:13-CV-1375-LMM, 2016 WL 3870080 (N.D. Ga. Feb. 17, 2016).

258. *Id.* at *1.

259. *Id.* at *2.

260. *Id.*

261. *Id.* at *7.

262. *Id.* at *11. *Cf.* *Kirksey v. Schindler Elevator Corp.*, Civ. Action 15-0115-WS-N, 2016 WL 5213928 (S.D. Ala. Sept. 21, 2016) (use of failure modes and effects analysis to determine inadequacy of design for escalator handrail).

263. No. 17C7567, 2020 WL 5570041 (N.D. Ill. Sept. 17, 2020).

264. 769 F. Supp. 2d 269 (S.D.N.Y. 2011).

something called Hazen Units. The defendant questioned whether the plaintiff's witnesses were "sufficiently expert in 'chemistry or phenol discoloration.'"²⁶⁵

Occasionally, the validity of a test rather than whether it was properly conducted may be the central issue. A case that involved toxicology testing rather than engineering best illustrates this point. In *Ruffin v. Shaw Industries, Inc.*, the plaintiffs claimed off-gassing from carpets was toxic.²⁶⁶ Their expert had developed her own test to determine toxicity, and the defendant challenged its validity. The trial court excluded her testimony, and the Fourth Circuit affirmed based on an application of the *Daubert* factors. The Environmental Protection Agency had attempted to replicate the test results, but could not do so.²⁶⁷

Conclusion

The two core messages of this reference guide are (1) the way in which engineering differs from science, and (2) the centrality of the design process and the evaluation of designs to the way engineers think and go about their work. Our review of cases in which engineering expert testimony is at issue shows that courts by and large admit testimony based on engineering design and analysis without requiring the kind of hypothesis testing and peer-reviewed publication that is central to the scientific method. When the testimony is more like science than engineering, however, at least some of the *Daubert* criteria for the evaluation of science are often applied; and sometimes admissibility depends primarily on an expert's knowledge and experience rather than either hypothesis testing or engineering analysis. These distinctions are fully in keeping with *Kumho Tire's* admonition that the objective in determining the admissibility of expert testimony is "to make certain that an expert, whether basing testimony upon professional studies or personal experience, employs in the courtroom the same level of intellectual rigor that characterizes the practice of an expert in the relevant field."²⁶⁸

265. *Id.* at 277.

266. 149 F.3d 294 (4th Cir. 1998).

267. *Id.* at 297–98.

268. *Kumho Tire Co., Ltd. v. Carmichael*, 526 U.S. 137, 152 (1999).

Glossary of Terms

artificial intelligence (AI). The ability of machines to make decisions and exhibit human-like behavior. AI combines mathematics, computer science, and data to augment human abilities.

complexity. The characteristic of a system being large, intricate, involving many parts, and displaying emergent behavior.

constraint. A fundamental or selected specification that limits the range of solutions.

design process. The collection of steps by which an engineer proceeds from a goal to a product, while making tradeoffs that satisfy given constraints.

engineer. A person who designs products, structures, computer algorithms, or the like, and/or oversees or participates in their implementation. Engineers also develop standards and tests for use in their work and do analyses of designs and the failure of designs to function as anticipated. An engineer may or may not have licensed credentials to practice. See professional engineer (PE).

engineering design. The process of devising a system, component, or process to meet desired needs.

error. In reference to statistics, the potential difference between a computed or measured value and the true value; more broadly, the difference between a predicted outcome and what actually occurs.

evidence. Within science and engineering, phenomena or data that tend to corroborate or confirm the validity of a hypothesis or a design or to show a hypothesis is wrong or that a design does not function as intended.

experiment. Testing methodology that involves intentionally manipulating some factor in a system to learn how that affects an outcome.

failure. The inability to perform according to the predicted or designed levels. Failure may be catastrophic if it results in the loss of lives, products, or property.

falsification. The idea that a scientific explanation is always contingent on some phenomenon or occurrence that shows it cannot be correct. Some accounts of how science works make falsification the touchstone of the scientific method, but more recent thinking makes it a less central factor.

hypothesis. Within science, a potential explanation for a specified set of natural phenomena. Note that this term is inconsistently applied outside of science.

licensure. See professional engineer (PE).

machine learning (ML). The subfield within the general artificial intelligence field that uses data and algorithms to learn, identify patterns, provide models, and make decisions with little or no human intervention.

model. Within engineering, usually, a mathematical representation of mechanisms and interactions within a system that can be used to investigate the impact of parameter changes on predicted system outcomes; or in some cases a physical device or structure used to approximate a design and facilitate its analysis.

prediction. Potential outcome of a scientific test or experiment that is arrived at by logically reasoning about what we would expect to observe if a hypothesis were true or false.

product. The ultimate outcome of an engineering design process.

professional engineer (PE). Designation earned after a graduate of an accredited engineering bachelor's degree program passes two tests and then obtains a state license.

science. A body of knowledge regarding the natural world and the process for building that knowledge based on evidence acquired through observation, experiment, and simulation.

scientific method. This method proceeds in a logical sequence from a statement of a question about why certain phenomena occur to communicating an explanation resulting from scientific inquiry. It is not a rigid lockstep process, but instead is dependent on publication, debate, and eventual consensus always subject to rejection or modification in light of new evidence.

structure. In the context of systems level models, something arranged in a definite pattern of organization.

systems engineering. An interdisciplinary field of engineering and engineering management that focuses on how to design, integrate, and manage complex systems over their life cycles. At its core, systems engineering utilizes systems thinking principles to organize this body of knowledge.

testing. The step in an engineering design or in a scientific pursuit where a product or a hypothesis is subjected to realistic operational conditions, whether in a laboratory setting or in its ultimate field of application.

theory. Within science, a concise, coherent, and predictive explanation for a broad range of natural phenomena that integrates and makes sense of many hypotheses. This term is inconsistently applied, and its use is more or less common in different disciplines and at different points in history. Hypothesis and theory are often used interchangeably, but a theory is more properly an explanation that has been accepted as correct, subject to modification or falsification.

tradeoffs. In the engineering design context, choices that necessitate striking a balance between two or more desirable but often competing criteria.

uncertainty. The idea that calculations or problems may not be precise, and that what actually occurs often can be predicted only within a defined range.

Reference Guide on Computer Science

BRIAN N. LEVINE, JOANNE PASQUARELLI, AND CLAY SHIELDS

Brian N. Levine, Ph.D., is Professor, Manning College of Information and Computer Science, University of Massachusetts Amherst.

Joanne Pasquarelli, Esq., Washington, D.C.

Clay Shields, Ph.D., is Professor, Department of Computer Science, Georgetown University.

CONTENTS

Introduction, 1412

Computer Organization, 1412

 Overview of Computer Hardware, 1413

 Binary Encoding, Bits, and Bytes, 1413

 Buses and Controllers, 1413

 Central Processing Units (CPUs), 1414

 Graphics Processing Units (GPUs), 1414

 Random Access Memory (RAM), 1415

 Data Storage, 1415

 Flash Memory, SSDs, and USB Drives, 1415

 Hard Drives, 1416

 Other Media, 1416

 General System Diagram, 1416

 Operating Systems, 1416

 File Systems, File Metadata, and Databases, 1417

 Virtual Machines, 1419

Algorithms, Programs, and Software, 1420

 Programs and Programming Languages, 1421

 Software Engineering, 1422

 Source Code Distribution and Software Licensing, 1424

 Source Code Review, 1424

 Reverse Engineering, 1426

Overview of Cryptography, 1427

 Encryption, 1427

 Shared Key Encryption, 1428

 Public Key Encryption, 1428

 Cryptographic signatures, 1429

 Certificates, 1429

Cryptographic Key Sizes and Attacking Encryption, 1430	
Attacking Encryption, 1431	
Cryptographic Hash Functions, 1432	
Attacking Cryptographic Hashes, 1433	
Cryptocurrency, 1433	
Networking, 1435	
Networking Layers, 1435	
IP Network and Link Layer Addresses, 1437	
Cloud Computing and Serverless Systems, 1439	
Wireless Networking and Mobile Devices, 1440	
Cellular Networks, 1441	
Wi-Fi, 1442	
Network Architectures, 1442	
Anonymity and Obfuscation, 1444	
Open-Access Wi-Fi, 1444	
Virtual Private Networking (VPN), 1444	
Anonymous Communication Systems, 1445	
Network Identifiers and Third-Party Data Collection, 1447	
Putting It All Together, 1450	
Investigating Network Traffic, 1451	
Investigating Peer-to-Peer File-Sharing Systems, 1452	
Investigating Anonymous Communication Systems, 1452	
Computer and Network Security, 1454	
Privacy, 1455	
Authentication, 1456	
What You Know (Passwords), 1457	
What You Have (Physical Tokens), 1458	
What You Are (Biometrics), 1458	
How Intrusions Happen, 1459	
Types of Attackers, 1459	
Types of Intrusions, 1460	
Legally Authorized Intrusions, 1461	
Malware, 1462	
Forensics and Digital Evidence, 1463	
Recovery of Digital Evidence, 1463	
Computer Science from the Perspective of Federal Rule	
of Evidence 702, 1465	
Experts in Computer Science, 1467	
The Quality of Forensics and Digital Evidence, 1468	
E-Discovery, 1469	
Glossary of Terms, 1470	

FIGURES

1. The components of a computer system, 1417
2. The IP network architecture is divided into several distinct layers, each providing services that build on the layers below, 1436
3. Network Address Translation box hides a larger network behind it from the rest of the internet, 1438
4. Racks of computers or “rack” servers in a data center, 1439
5. A client-server network architecture. Dotted lines represent connections across the internet, 1443
6. A peer-to-peer network architecture. Dotted lines represent connections across the internet, 1443
7. Tor Browser connects a user to a website in a multi-proxy setting, 1446
8. Tor onion services connect a user to a website in a multi-proxy setting. Unlike Tor Browser, the IP address of the website is unknown to the user, 1447
9. Left: A user’s view of the file system interface. Right: The true lifecycle of data and files in a file system, 1464

Introduction

Computers have become inescapable. Virtually every adult and teen carries and uses a computer in the form of a smartphone holding incredible amounts of information about the user. Computers are present as wearable devices that monitor our health or are present in our buildings as fixtures that monitor and control our environment. Cars and critical infrastructure no longer work without them. We commonly use computers to communicate, seek and find information, navigate, and supplement our memory—and evidence of our actions often remains. Companies store massive amounts of data electronically about their activities and, where applicable, those of their customers. Accordingly, our everyday movements, communications, and actions often result in critical evidence that is relevant to both criminal and civil disputes. Court cases often depend on a forensic analysis of this stored digital evidence of the computational activity that has occurred or been recorded, particularly in the realm of emails and other communications. Computers will be part of many legal cases, in some aspect, for the foreseeable future.

In this reference guide, we provide a brief overview of how computers operate and communicate; the basics of algorithms, software, and cryptography; how computers communicate over networks like the internet; computer security goals and the process of meeting them; and what digital forensic evidence can be collected to support or refute some legal theories. Given the wide scope of topics, we are necessarily brief in our treatment of these topics, but we hope this reference guide provides a basis for understanding the technology that is involved in many legal issues. Standard textbooks are available that cover many of these topics in greater detail, such as from Kurose and Ross,¹ Anderson,² Bishop,³ and Kernighan.⁴ See also Darrell⁵ and Kerr⁶ for legal textbooks concerned with technology and computers.

Computer Organization

We begin with a general overview of computer organization, including hardware, file systems, databases, and virtual machines. Computer networking is covered in a later section of this guide.

1. James Kurose & Keith Ross, *Computer Networking: A Top-Down Approach* (7th ed. 2016).

2. Ross Anderson, *Security Engineering: A Guide to Building Dependable Distributed Systems* (3d ed. 2020).

3. Matt Bishop, *Computer Security: Art and Science* (2d ed. 2018).

4. Brian W. Kernighan, *Understanding the Digital World: What You Need to Know about Computers, the Internet, Privacy, and Security* (2d ed. 2021).

5. Keith B. Darrell, *Issues in Internet Law: Society, Technology, and the Law* (11th ed. 2018).

6. Orin Kerr, *Computer Crime Law* (5th ed. 2022).

Overview of Computer Hardware

A computer is a tool that processes information at blazing speeds. It performs this feat by encoding the information as electrical signals that can be processed as billions of operations per second, stored in huge quantities, or sent over a network. The hardware of a computer is designed to process data quickly through simple mathematical operations. It is the combination of computational power with functionality provided by programs that makes it incredibly useful. We describe programs and how they are created in “Algorithms, Programs, and Software.”

Binary Encoding, Bits, and Bytes

Every computer performs computations using a series of high and low voltages that represent a series of 1 or 0 values, each value being called a bit. All information on the computer, including programs or data files, is a series of bits, with the series representing some numeric value that has meaning in a particular context. This representation is called binary encoding. Each added bit in a series doubles the number of possible values that can be represented. A single bit can represent two values; two bits can represent four values; three bits eight values, and so on. The most common unit of data is 8 bits combined; this unit is named a byte, and it can represent 256 possible values and often represents one character of text. Most data are measured in terms of bytes, and it is common for measurements of bits to be represented with a lowercase b and bytes with an uppercase B. Units of a thousand bytes are called a kilobyte, abbreviated KB; units of a million are a megabyte, or MB; units of a billion bytes are gigabytes, or GB; and units of a trillion bytes are terabytes, or TB.

Buses and Controllers

Almost all computers have a similar physical structure. Typically, there is a motherboard that provides a base for other components and for communication between them, using electrical connections called buses. The buses connect components called controllers that handle receiving data from the user, from storage, or from the network and sending that data to the central processing unit (CPU). A typical motherboard will have a variety of buses, some that are very high speed to access random access memory (RAM) or to other components like graphical processing units (GPUs), and others that are lower speed and connect devices like Universal Serial Bus (USB) mice and keyboards.

Central Processing Units (CPUs)

The primary computational component that is mounted on the motherboard is the central processing unit (CPU), which is often the single most expensive component. Older computers most often had one CPU that contained one computational core whose speed was governed by a clock that generated electrical signals to keep computations internally synchronized. Some higher-power computers had more than one CPU mounted on the motherboard and could perform multiple computations at the same time with one set of computations in each CPU.

Advances in computational speed initially came from increasing the tick speed of the internal clock. However, this approach reached physical limits, mostly in terms of power consumption and related heat generation. More recently, designers started adding additional computational cores to a single CPU so that each CPU could perform multiple computations simultaneously. Now a single physical CPU typically has many computational cores, which enable new uses like hosting virtual machines, as discussed below. The number of cores varies depending on design and cost of the CPU; low-cost CPUs might have 1 to 4 cores, while high-end CPUs might have 64 cores or more, with a range in between. With mobile computing becoming widespread, designers often work to decrease the power required for computation so that mobile devices, like phones and laptops, can last longer while running on battery power.

Graphics Processing Units (GPUs)

The other component that is increasingly used for computation is the graphics processing unit, or GPU. The GPU was originally used to increase the speed of on-screen video, particularly for games, but over time other uses have become popular. GPUs are excellent for machine learning and cryptocurrencies, and they have become essential to those areas.⁷

GPUs differ from CPUs in that they have many more computational cores than a CPU, though each core is more specialized and limited in what it can do. Whereas a CPU might have between 4 and 32 cores, a GPU might have thousands of less-powerful cores. In addition, GPUs execute tasks delegated by the CPU during the execution of a program, and they do not run programs on their own.

7. For a detailed discussion of machine learning, see James E. Baker and Laurie N. Hobart, *Reference Guide on Artificial Intelligence*, “What is Machine Learning?” section, in this manual.

Random Access Memory (RAM)

The CPU stores some working data in its own circuitry, but such storage is expensive and displaces circuitry that could be used for computation. Each CPU has several caches of memory: the high speed but smaller level 1 cache, and one or more other levels of caches, each larger and slower than the prior level. More storage is supplied off the CPU in random access memory (or RAM). RAM is a large bank of hardware memory that can be accessed relatively quickly by the CPU when it needs data. RAM is often connected to the CPU on a dedicated, high-speed bus so there is no conflict with other data coming from other devices. GPUs have less cache, instead having high-bandwidth connections to the RAM to supply the data needed.

RAM is volatile memory, meaning it requires continuous power to store data. For desktop computers this means that data can be lost when power is lost. For laptops and some battery-powered devices there is usually a mechanism to store the contents of RAM in longer-term storage when it appears the battery will become exhausted; other devices have a memory-preserving sleep mode.

Data Storage

Flash Memory, SSDs, and USB Drives

Most computers now use flash memory for internal data storage. Flash memory is a type of hardware-based storage similar to RAM, except that it is nonvolatile. When power is removed from flash memory the data is not lost. Flash is not a replacement for RAM in part because it is much slower and in part because flash has a characteristic that the memory cells can only be written to a limited number of times. Flash memory is commonly used for solid state drives (SSDs). These drives are an improvement over hard drives that are based on spinning magnetic platters, described below. SSDs have faster access speeds and lower power usage, but they can be more expensive for the same amount of storage.

Flash is also commonly used in portable storage. It is incorporated into a huge range of cards, sticks, and keys, which sport a variety of connectors. Perhaps the most common type of portable storage is the USB drive—often referred to as a “USB key fob” or “USB key”—which connects via a computer’s USB port. Card readers that support a variety of formats are available, and some laptops include built-in readers.

Hard Drives

Non-SSD hard drives use an older mechanical technology based on magnetism to store bits of data on a rotating platter. While these drives are less common in laptops and PCs, they are still used in many server applications for bulk data storage. The size of the platter has been getting smaller over time, though the density of information written on the platter continues to increase. Desktop drives commonly use 3.5-inch platters, while laptops use 2.5-inch platters, and embedded devices and the smallest laptops use drives with 1.5-inch platters.

Other Media

While SSDs and hard drives are most common, there are other types of media still in use. Optical media, such as compact discs (CDs), digital video discs (DVDs), and Blu-ray discs, encode data in a way that they are readable by laser light. Current technologies store data on discs by encoding bits as tiny pits that change the reflectivity in reflective media and can be read as bits by a laser. Tape is sometimes used as a backup media. It is a long, magnetic, ribbon-like material that is stored on spools that are often inside a cartridge. Tape provides sequential access to data instead of random access. The tape must be read from front to back. Tape can provide significant amounts of storage inexpensively, though the access rate is slow.

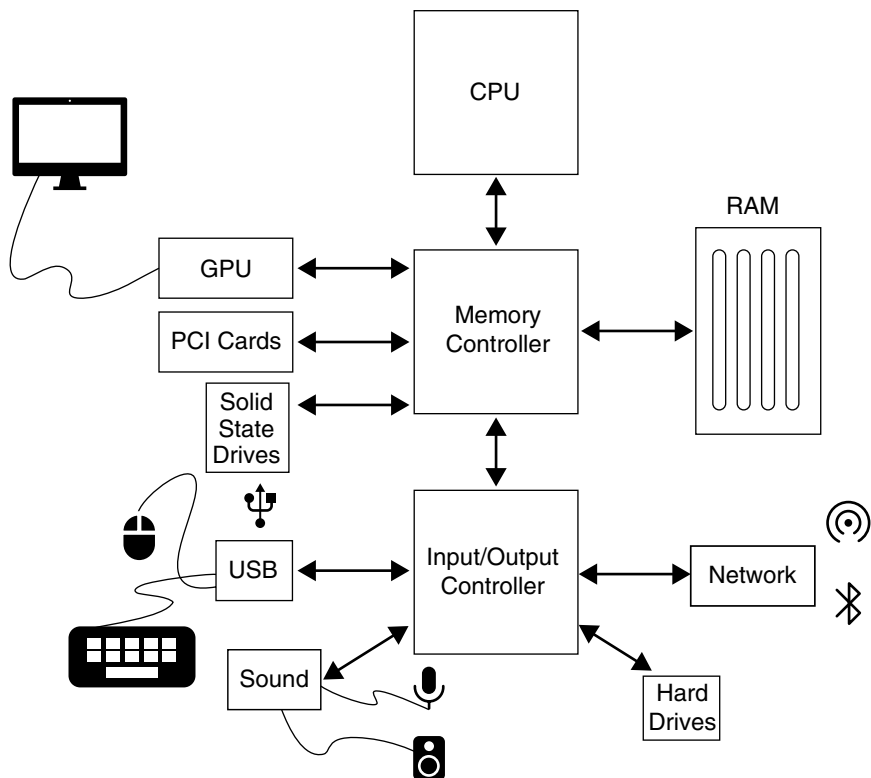
General System Diagram

Figure 1 shows the basic layout of the devices in a computer. The high-speed bus and low-speed bus are separate components on older computers, but some or all of that functionality might be incorporated directly into modern CPUs.

Operating Systems

The hardware of a computer is shared across all the programs using it. Operating systems (OS) are a common and specialized type of software that mediates between the system hardware and the software that wants to use it. Operating systems provide a number of useful functions. They manage resources shared between programs, such as access to devices like storage and input/output devices and allocating memory. They provide a common interface to hardware so that programmers don't need to be aware of specific differences between individual hardware brands or components by providing an interface for drivers, which specify how the OS can interact with hardware. They provide a user interface so that users can start

Figure 1. The components of a computer system.



and stop programs and access files and other information. Computer programs written for one operating system are generally not compatible and will not work with others. Some OS are designed for desktop and laptop use, like Microsoft Windows and Apple's macOS. Others are used for mobile devices, like Google's Android or Apple's iOS. Still others are most commonly used for servers, like Linux.

File Systems, File Metadata, and Databases

Computer storage media provide the ability to store information, but they do not inherently organize it; they come as large blank storage areas. Computer operating systems format storage media with a file system that provides a structure for storing and retrieving directories and files. Operating systems typically protect

files that are critical for the OS to function and isolate user-generated files between users. The file system also keeps accounting information about each file. This is generally referred to as the MAC times, which are the times a file was last modified, when it was last accessed, or when it was created. These times may or may not be entirely accurate, depending on the operating system being used, if the system clock is accurately set, and the circumstances of the file creation; they also might be manually changed. They are not inherently reliable. These data are a type of metadata, which is data about or related to other data.

File metadata can be instrumental in litigation as it can reveal many facts relevant to key issues in litigation, such as the location and the date and time files were created, accessed, or modified. Often the most important role of file metadata is to support or attack the credibility of evidence—both testimonial evidence and document evidence. For example, in a contract dispute metadata revealed the creation date of a key document was almost one year later than the defendant had represented. Of note, the creation date shown by the file's metadata was in fact just two days prior to the document being disclosed to the court. While metadata still requires authentication to be introduced as evidence of the true creation date, it provided sufficient grounds to challenge the credibility of the defendant's prior assertions regarding the existence of the document.⁸ A court may deem metadata to be self-authenticating if the opposing party shows no reason to doubt the authenticity of the metadata.⁹ Metadata regarding the time a file is created, accessed, or modified can also be relevant in medical malpractice cases where electronic medical record systems are utilized. In today's world of electronic files, metadata can be relevant in any case where the creation and access times for key files are important to pending litigation.

When a file is written to storage by a program, it is itself formatted in a very specific way according to the design of the program and the information the program needs to store. The internal format of the file depends on the program being used, but many files include additional useful information about the file that may not be visible to the user without special tools or without using program features to access it. For example, one image file format is the exchangeable image file format (EXIF) standard, which can include the time and date the photo was taken, a GPS location, as well as details on the camera that took the image. Similarly, Microsoft Office documents contain metadata about the creation of the document, often including the author and the amount of time spent editing the document and changes made, among many other things. Any file under the control of a user can be altered. Thus, just like the images themselves, EXIF data could be easily altered by the owner of a file, meaning the authenticity of the metadata should be confirmed before accepting it as evidence.

8. *SPV-LS v. Transamerica Life Ins. Co.*, 912 F.3d 1106, 1114 (8th Cir. 2019).

9. *Tamares Las Vegas Props., LLC v. Travelers Indemnity Co.*, 586 F. Supp. 2d 1022 (D. Nev. 2022), *cited by* *CNA Ins. Co. Ltd. v. Expeditors Int'l of Wash.*, 2023 WL 6892565 (W.D. Wash.).

EXIF data can become critical evidence in criminal cases. For example, an image of child sexual abuse material (otherwise known as child pornography) containing EXIF data revealing the GPS location where the image was taken with an iPhone 4 was posted to a website. This EXIF data led the FBI to the owner of the iPhone who admitted to taking the image. During the resulting prosecution, the defendant filed a suppression motion arguing the EXIF data was protected by the Fourth Amendment because it was not visible to anyone viewing the image on the website. The court denied the motion holding there is no Fourth Amendment protection for the EXIF data as it was publicly accessible when the image was publicly posted.¹⁰

It is also important to note that the internal format of the data may not translate directly to what is shown on the screen. Additional information may be recoverable from careful analysis. For example, a portable document format (PDF) file is used to format and display documents for printing. Making a PDF file is often the final step in releasing documents in public settings. Information that was intended to be redacted from PDF documents is sometimes recoverable with special tools or simple mechanisms. For example, if the redaction involves covering text with a black rectangle, the redacted text may still be stored within the file and recoverable using the right tools.

A database is an alternative way of organizing a collection of information. Instead of files, the database has entries stored in tables. Most often, all items in a table are related information. For example, one table might consist of a set of names; another table might consist of a set of addresses. Internal identifiers are used to relate items across tables. Databases use a different interface than file systems. Most often, there is a computer language that can issue queries to the database so that it can retrieve items that match the queries. (For example, a query may be issued to identify all rows in a table related to a specific city.) In contrast, files are most often accessed by the file name and perhaps what directory the file is in; while some operating systems support searching files by content, it is not the primary means of accessing data.

Virtual Machines

Computers exist in the physical world, each taking up some space and requiring cabling for power and networking—which can be inconvenient for those who need an additional computer, for example, to run different operating systems or for those who want their own isolated server for security reasons. Providing these individual devices—real, physical devices—can be cumbersome and expensive.

10. *United States v. Post*, 997 F. Supp. 2d 602 (S.D. Tex. 2014).

Virtual machines (VMs) have become a commonplace solution to this problem. A virtual machine is a software program that pretends to be computer hardware while running on a host's real hardware. A virtual machine makes use of one or more of the host computer's CPUs or computational cores like any other program running on the host. The VM is allocated storage on the host computer, and it provides network access through the host computer. Within the environment provided by the VM, it is then possible to install the same or a different operating system as the host and then run programs on the VM. Programs operating in a virtual machine most often run the same way they do on a physical machine; some malicious software will attempt to detect if it is in a virtual machine and change its operation as a result, but few if any benign programs do so. Virtual machines are designed to be completely separate from each other and to isolate operations in the virtual machine from the host systems and other VMs.

Virtual machines have several uses. The most common is internet server hosting. Instead of paying to have an expensive physical server placed in a secure co-location facility, users and companies can instead rent virtual machines from services such as Amazon Web Services, Google Cloud, or Microsoft Azure, among many others. The use of rented servers is often called cloud computing or a serverless architecture (see section titled "Cloud Computing and Serverless Systems" below) and represents a multibillion-dollar market, although not all such services are provided through virtual machines. VMs are also commonly used to increase security. Untrusted software can be run in isolation on a VM in case the software acts maliciously. If it does, the impact is isolated, and the virtual machine can be easily discarded. Finally, individual users might run virtual machines to experiment with different operating systems. Rather than having a computer for each, one computer can host many different operating systems that can run different software.

Algorithms, Programs, and Software

The operation of programs is central to understanding other topics surrounding computing. Here, we provide an overview of how programs are developed and turned into software products, which is relevant to legal issues that address what software does and how it was developed.

An algorithm is a series of ordered steps used to complete a task.¹¹ We use algorithms in many aspects of our daily life: a recipe to cook a certain dish or the diagnosis and repair of an item are both common examples of algorithms that

11. Algorithms (and in particular complex algorithms) are discussed in detail in James E. Baker and Laurie N. Hobart, *Reference Guide on Artificial Intelligence*, in this manual. See, e.g., sections titled "Complex Algorithms," "The Heart of AI Is the Algorithm," "Forms of Algorithmic Bias," and "Algorithms that Predict Human Behavior."

humans follow. An algorithm must be specified in terms that the person following it can understand and complete; you couldn't give a child direction on how to drive somewhere as they don't know how to drive, for example. It is also possible to define an algorithm as a combination of many smaller algorithms. For example, one might follow one algorithm to diagnose the cause of an item's malfunction, and then repair the item following an algorithm determined by the diagnosis.

Programs and Programming Languages

A computer program is an algorithm that a computer can run to solve a computational problem or task. An application is a program that a user interacts with directly. When a program is in a format that the computer's hardware can execute directly, it is called an executable. Because this executable format is difficult for humans to edit, programs are instead created by programmers who determine what the algorithm should do and then write source code that specifies the steps of the algorithm in a human-understandable programming language.

Some programming languages are called interpreted languages; sometimes these are called scripting languages, or scripts. In these languages there exists an executable called an interpreter that takes the source code and executes the algorithm one step at a time, essentially performing the interpretation and execution concurrently. Other languages are compiled languages. Programs in these languages are translated into executable form using a program called a compiler. This resulting executable can then be distributed and run independently of the source code. For compiled languages, translation to an executable occurs entirely ahead of the execution of the program. In general, compiled programs have better performance than interpreted programs but require more programmer time to develop and are closely tied to the system where they are expected to run; they often cannot run on other systems without changes or special support. To be able to run on many different types of hardware, some languages are designed to use an intermediate format that is independent of the hardware being used and which gets translated as needed to run on a local processor. For example, Java is designed to run on a Java virtual machine (JVM) or equivalent. Instead of recompiling the Java source code for each computer platform, the JVM only is recompiled. To execute a compiled Java program, the JVM translates from the program's instructions to the actual hardware instructions; performance is better than purely interpreted code.

There are many different programming languages, each of which is generally used to solve some class of problems. Some languages are most convenient to use for the web, with some on the web server (such as Ruby and PHP) and some in the web browser (like JavaScript); some for data analysis (like Python, R, and Julia); others for games (commonly C++ and C#); others for computational infrastructure (like C and Rust); and still others for mobile application

development (like Swift and Java). There is a significant overlap between languages, and most can be used for multiple purposes.

It is also possible to partially reverse the compilation process, which is called decompilation. It is not straightforward to decompile code. For example, decompilation typically does not produce the exact source code used to create the executable because information of use only to human programmers, like names or labels indicating what a value means, is often removed as a step in the compilation process.

Computer programmers often split programs into smaller pieces for a variety of reasons, including to reuse algorithms and to allow many programmers to work on the same program at once. Because of this, code is often written in a modular form where smaller sub-algorithms can be assembled to form the larger overall algorithm. Within a program, these sub-algorithms might be called functions, methods, or procedures interchangeably. Often, so that such functions can be shared among many programs, they are placed into libraries. These libraries might natively be part of the ecosystem of the programming language; they might be libraries that are freely available for download; they might be libraries that are commercially available for license; or they might be part of the computer's installed operating system, as described above. The term software refers to any or all of these various executable portions of code, from applications to libraries to the operating system.

Software Engineering

Programming is just one aspect of the larger topic of software engineering. Software engineering is a process that also involves design, testing, documentation, and maintenance of software. Many factors affect the complexity of the process, including the number of users of the software, the complexity of the software, the longevity of the software, the level of reliability needed, and the number of programmers coordinating simultaneously or longitudinally over time. A critical aspect of software engineering is communicating and documenting the requirements of a project.¹²

A variety of tools are commonly used to manage these processes. For example, splitting software into modules helps with the development of larger projects. Different teams of programmers can work on and test each module separately,

12. For a general discussion of the engineering design process, see Chaouki T. Abdallah et al., *Reference Guide on Engineering*, “The Engineering Design Process” section, in this manual. The role of computers and the implications of their use in design and complex systems is discussed in the *Reference Guide on Engineering* in the “Complex Systems” and “Computers, Artificial Intelligence, and Machine Learning” sections.

then combine them into a larger program. This approach leads to a model where most programs are layered. Each layer is independent and presents an interface known as the application programming interface (API) that defines how the software in that layer can be accessed. Each layer can then be changed or replaced without major modifications overall as long as the layer maintains a consistent API. APIs can be an important part of commercial products. For example, services like Amazon Web Services, Google Cloud, and Microsoft Azure provide computational services for rent that can be accessed through an API.

Software projects can vary significantly in size. Software size is often measured in lines of code, which is what it seems: a count of how long the programming language portion is. This metric is not easily comparable between languages and is a rough approximation of effort at best.

A small module or program might be only a few tens or hundreds of lines of code. A large project, like the source code for a computer operating system, might contain thousands of smaller modules and be many millions of lines long. It is thought that the code for Microsoft Windows totals over 50 million lines; the Linux operating system has about 28 million lines of code as of the time of this writing.

Code is typically stored in a source code repository. This is a specialized database, often called version control software, that can track changes made by different programmers to the code, which can help programmers and forensic investigators alike understand when and where code was added. A repository is helpful in merging new code with existing code, including new code in a program for testing, and submitting code for managerial or peer review before final inclusion. It also allows for different code versions and reverting source code changes should there be errors. Common tools for source code management include git, CVS, svn, and Visual SourceSafe. These can be useful sources that can show how software was developed over time.

Many kinds of testing methodologies are involved in verifying the correct operation of software. For example, unit tests ensure that the smallest components of a code base function as expected. Regression tests ensure that as new features are added or as repairs are made the current level of performance of a software system is not reduced or new errors added. Integration tests ensure that distinct subsystems interoperate as expected. Often software systems involve a continuous integration continuous delivery (CI/CD) pipeline where new changes are tested against a battery of tests before they are accepted into the main code base; this testing often occurs overnight. The goal of CI/CD is that software is always available in a state that is operational and correct, so that it can be delivered to the customer at any time. As a code base is expanded, developers add to the battery of tests to keep the CI/CD pipeline current. Finally, many larger projects will involve a team of quality assurance (QA) tests that test for bugs, rate performance, and ensure a high-quality user experience before software is released.

Source Code Distribution and Software Licensing

Most commercial programs that are developed by companies do not disclose source code, and the program's source code is not shared with others. This closed-source approach helps protect trade secrets and avoids aiding competitors. In contrast, some programs are developed openly, and the source code is publicly available. For example, Microsoft's GitHub.com platform hosts many open-source projects.

Software is commonly licensed, rather than being sold outright. Licensing allows the software vendor to impose conditions on its use, including prohibitions on copying, distribution, and reverse engineering. Many commercial programs are licensed under proprietary End User License Agreements (EULAs) that users must accept to acquire or use the software, and copyright is retained by the company.

Open-source software is often distributed under what are commonly called Free and Open Source (FOSS) licenses and are sometimes referred to less accurately as "free software." There are a variety of open-source licenses; most allow use, copying, and modifications of the program. They vary on whether the code can be sold, redistributed, or sublicensed. Some licenses, most notably the GNU Public License (GPL), allow modification and private use but require that any such modifications that are sold or distributed are also licensed under the GPL, ensuring that source code improvements become available and that software will be improved over time by interested parties but remain available to all. As this approach is quite different from usual copyright, this license is referred to as copyleft.

Other open-source licenses allow sale and modification without requiring source code distribution; the BSD license, for example, allows use of and modification of code if the original license is included in future distributions. Similar licenses exist for other material, such as writing and photographs. A common one is called the Creative Commons license.

Source Code Review

In many situations, parties in a legal proceeding disagree about what software does, making source code review a key part of the discovery process so that both parties can assess essential evidence. This scenario arises in cases involving product liability, trade secrets, and patent infringement and is sought by defendants in criminal cases where key evidence was collected through software. Source code review intrudes on information protected as proprietary, trade secret, or law-enforcement sensitive information and therefore should only be used when the threshold has been met to overcome these software protections. Source code

review occurs when the software code is the central evidence to the case, such as in the seminal software copyright case of *Google, L.L.C. v. Oracle America, Inc.* Source code review was also used by plaintiffs in a key Toyota product liability case to argue that the software governing the electronic throttle control system was defective, causing sudden acceleration events that resulted in numerous deaths and injuries.¹³

However, protections for trade secret, proprietary, and law enforcement sensitive information should prevent disclosure of source code when the software is not the central evidence of the case itself. For example, in a patent infringement case involving wireless communication technology, a U.S. district court denied the plaintiff the opportunity to review a third party's source code, holding that the third party's concerns regarding the security of the source code could not be ignored despite the protective order. The court stated that the entity owning the source code was not a party to the lawsuit nor were they accused of patent infringement. Furthermore, the court held that the relevant information could be obtained from deposing a company engineer rather than from reviewing the source code.¹⁴ Similarly, software used in criminal investigations becomes essential when it is used to collect evidence in the prosecution's case-in-chief. For example, in *United States v. Budziak*, the Ninth Circuit held that the source code was material to the defense's case relying on the fact that the evidence presented at trial to support the distribution of child pornography charge was devoted to the government's investigative software (called eP2P) and the evidence that was collected by its use. No court has held the investigative source code was material to the defense when the prosecution's case-in-chief does not include evidence collected through the software.¹⁵

The demand for source code review in civil cases has resulted in an industry of source code review companies offering experts who can conduct code reviews and testify in court proceedings. Often the code being reviewed is proprietary and may contain information like trade secrets, so the source code is made available only on a secure machine, disconnected from any network, sometimes at one of the parties' representatives' offices. The parties often agree to provide the same set of tools to be used in the code review that the programmers used to develop it, or sometimes an even more minimal subset of similar tools is chosen by the legal representatives without expert guidance.

Programmers need to change code, however, to create new features and then turn those into a working, executable program. Source code reviewers usually

13. *Bookout et al. v. Toyota Motor Sales USA Inc. et al.*, No. CJ-2008-7969, *verdict returned* (Okla. Dist. Ct., Okla. Cty. Oct. 25, 2013).

14. *Realtime Data L.L.C. v. MetroPCS Tex. LLC*, No. 12cv1048-BTM (MDD) (S.D. Cal. May 25, 2012).

15. See *United States v. Pirosko*, 787 F.3d 358 (6th Cir. 2015); *United States v. Hoeffener*, 950 F.3d 1037 (8th Cir. 2020); *United States v. Arumugam*, 2020 WL 949937 (W.D. Wash.); *United States v. Feldman*, 2015 WL 248006 (E.D. Wis.).

have different needs, as their goal most often is not to create or redesign code. Instead, the focus is on being able to read existing code easily to understand the underlying algorithm, to search the code for features, and to print code for reference in a way that allows ease of reference in legal discussions. This goal requires a different set of tools that focus on reading and searching code. For example, *doxygen* is a tool that converts source code to non-editable local files that can be opened as web pages that include the file name and line numbers for easy reference to the original source, and *dnngrep* is a free tool that can conduct fast searches.

It is worth noting that the term “source code” might refer to different things depending on context. To a computer scientist, the term typically refers to the programmatic source code that is compiled or interpreted to produce a program. In a legal context, however, the term commonly applies to all files that are produced as part of discovery, which can include many other related files. These often include the following: build system configurations files or scripts, which are used to compile and link together the many modules that make up a large program; code that conducts testing of the program to detect errors; and other files, such as design documents, notes about the code, or histories of when files were added during development. These additional files often provide useful context to a source code reviewer.

Reverse Engineering

Aspects of hardware and code that are protected as a trade secret are subject to reverse engineering, in which others attempt to discover the secrets through independent discovery. *Kewanee Oil Co. v. Bicron Corp.*¹⁶ established that trade secrets do not receive patent protection and that it is legal to reverse engineer items in the public domain.

Reverse engineering is also used as an innovative tool for chip design. In 1985, Congress passed the Semiconductor Chip Protection Act (SCPA),¹⁷ which protects semiconductor design while providing a reverse engineering provision that allows copying of the entire chip design. The law does not distinguish between the protectable and nonprotectable portions of the chip as long as the copying is for the purpose of teaching, evaluating, or analyzing the chip. The law was written to specifically protect industry practice and encourage innovation. In intellectual property litigation where chips appeared to be similar, Congress assumed that admitting expert testimony to assist in determining whether subtle changes in a chip layout (called a “mask work”) were significant would resolve the problem of distinguishing a copy from a legitimate reverse-engineering attempt in most cases.¹⁸

16. 416 U.S. 470, 94 S. Ct. 1879 (1974).

17. 17 U.S.C. §§ 901–904 (1984).

18. *Altera Corp. v. Clear Logic, Inc.*, 424 F.3d 1079 (9th Cir. 2005).

Reverse engineering of software is often prohibited. License agreements and non disclosure agreements can limit the right to work with software to determine its function. A copyright owner can disallow making copies of a work for reverse-engineering purposes. The Digital Millennium Copyright Act (DMCA) prohibits “circumvention of ‘technological protection measures’ that ‘effectively control access’ to copyrighted works.” In some cases, however, it has been found legal to reverse engineer software in particular circumstances, primarily to provide interoperability between systems.¹⁹

Overview of Cryptography

With a basic background of computing hardware and software complete, we move toward a brief description of the cryptographic mechanisms that have become essential tools for securing information stored on computers and transmitted across networks. These algorithms are fundamental to many areas of computer science as they relate to legal issues. We introduce concepts here that appear in later sections.

Encryption is used to maintain the secrecy of information. Encryption can be a legal issue when, for example, critical evidence is held inaccessible within an encrypted file. Cryptographic hashing algorithms are widely used to preserve the integrity of data and to recognize known content accurately and easily. Cryptocurrencies are complex networked systems that can allow transfers over the internet of funds that have public value without the involvement of financial institutions; their rise has enabled new types of criminal activity and new ways to launder money.

Encryption

The most common way to ensure the confidentiality of data is to encrypt it. The net effect of encryption is to take a large secret (the data being encrypted) and reduce it to a small secret (the key used for encryption). This approach makes handling sensitive information easier, as data can be stored and transported securely on the open internet in an encrypted form. In general, if the data is obtained by others, it will remain confidential as long as the key remains a secret. The keys are transported by a separate, more secure mechanism, which is easier because they remain small regardless of the size of the data. The terms passwords and keys are similar concepts but often used differently. The term password most often refers to a collection of characters that can be entered with a keyboard (letters, numbers, and punctuation) and are human-readable. A key is a more general term as the collection can contain additional values that cannot be entered

19. *Sega Enters. v. Accolade, Inc.*, 977 F.2d 1510 (9th Cir. 1992); *Sony Computer Ent. v. Connectix*, 203 F.3d 596 (9th Cir. 2000).

with a keyboard, or perhaps a collection of values that is too long for a human to reasonably enter. In other words, the set of possible keys is much larger than the set of possible passwords. Keys are best chosen randomly, but they can be derived from passwords when needed so a human can remember them more easily.

When we encrypt data, we are altering data so that it no longer has any visible structure. The process of alteration is governed by the encryption algorithm and the encryption key chosen. Taking the original bits, referred to as the plaintext, and running them through the encryption algorithm with a given key produces the ciphertext. A good encryption algorithm produces a ciphertext that appears to have been generated completely at random. When we later want to recover the plaintext, we decrypt it by running the ciphertext through the encryption algorithm with the correct key.

If the algorithm works correctly, the secrecy of the data lies entirely with the key. For proven encryption algorithms the operations are publicly available and the best-known attack is to painstakingly attempt every possible key. We say that information is secure if an adversary cannot possibly attempt all keys in any reasonable amount of time given the resources (i.e., time and money) they have available or before the information loses its value. The number of possible keys for modern algorithms is so large that it is impossible in practice to search through all of them for the correct one, as we discuss below.

Shared Key Encryption

Algorithms that use the same key for encryption and decryption are called shared key or symmetric key algorithms. The standard shared key algorithm today is the Advanced Encryption Standard (AES), which is commonly implemented in hardware on many modern computer CPUs to increase performance.

While shared key encryption is common and efficient, it does have its drawbacks. The primary one is that if you want to have secure communication with many independent people, you need to share a different key with each individual separately. A simpler approach to this key management problem is to use public key encryption, as we describe below. The advantage of shared key encryption algorithms is that they are generally much faster than public key algorithms.

Public Key Encryption

Public key encryption algorithms use a linked pair of keys: anything encrypted with one key can only be decrypted using the other key. It is called public key encryption because one of these keys—the public key—is not kept secret. The other key is the private key and is kept secret by the owner. The advantage of this approach is that everyone who wants to send a confidential message to a recipient can

use the recipient's public key for encryption; pairwise keys between individuals are not required as they are in shared key algorithms. All messages can be decrypted by the recipient using their private key.

Public key algorithms are based on the fact that mathematically there are problems that are difficult to solve but very easy to check if a solution is correct. By analogy, it is easy to see that a jigsaw puzzle has been solved; but it takes effort to complete a puzzle given a bag of pieces. A real cryptographic example is that given two large prime numbers it is easy to multiply them together to determine if their product is equal to a third value. But given a large number alone, it is difficult to factor out the original prime numbers. One of the original public-key algorithms, called RSA after the initials of its inventors, was based on this factoring challenge. Other algorithms are based on similar mathematical problems that are easy to solve for the person who has the private key but essentially impossible to solve without it.

Because these algorithms work differently than shared key algorithms, the key size is not an accurate indicator as to how long it would take to search for and find an unknown key, and because of the mathematical operations that need to be performed, public key encryption is often much slower than shared key encryption. It is also possible that some breakthrough in mathematics or computer science will render a previously difficult to solve problem suddenly easy to solve, making any algorithm based on that problem immediately breakable. Common public-key algorithms include RSA, elliptic curves, and El Gamal.

Cryptographic signatures

Public key encryption can also provide proof of who sent a particular message, called a digital signature. The basic idea is that if someone encrypts some data with their private key, decrypting it with the known public key verifies it was encrypted by the private key owner. Proving it was encrypted with the private key is therefore equivalent to the private key holder signing it (under the assumption that the private key would never be disclosed by the owner). For convenience and speed, some digital signatures do not encrypt the entire document. Instead, a cryptographic summary of the document is computed (called a hash, as described below), and that is encrypted. The hash is small and unique to the document being signed. Not all cryptographic digital signatures operate this way, but it's a good high-level model for understanding how digital signatures are computed.

Certificates

One problem with public key cryptography, however, is verifying who owns a particular key. This verification must be managed by computers, be performed

in milliseconds, and be mathematically based. The common solution to this is the use of cryptographic certificates, which make use of digital signatures to verify public key ownership; certificates provide much of the cryptographic foundation of the internet. Individual operating systems or software, such as web browsers, are preloaded with root certificates. These certificates are, in short, the public key of trusted organizations that serve to verify other certificates. In this hierarchical system of trust, a shopping website would create a public key to include it in a certificate of its own, and then pay one of the root certificate organizations to sign it. Consumers are trusting the root certificate organizations to sign with integrity.

For example, when a secure connection is made over the internet to shop using a web browser, the shopping website presents its own certificate to the browser. As described above, this certificate contains the public key for the shopping website and a digital signature from one of the root certificate organizations. The browser uses one of the preloaded root certificates to verify the trusted organization's signature on the shopping website certificate. When the browser finds that the shopping site certificate is correctly signed, it then trusts that the shopping site certificate belongs to the shopping site (and not some impostor) and that it is safe to conduct transactions. The site public key is then used to exchange a shared key used during only this session for encrypting the transactions between the user's browser and the shopping website. The actual process has a few more steps, including some intermediate certificates and keys, but conceptually the result is the same.

Cryptographic Key Sizes and Attacking Encryption

Cryptographic operations depend on the fact that the number of possible keys and hashes are so large that they represent values far outside human experience. Each key or hash is a series of individual bits, each a single 1 or 0 value, often hundreds of bits long. Each time a bit is added, the number of possible combinations for the series doubles. By the time the length of the series is 128 bits, the number of possible combinations is 2^{128} which is about 10^{38} (or 1 followed by 38 zeros). To put this in perspective, some estimates put the number of grains of sand on the earth at about 10^{24} . In other words, there are about 100 trillion times as many possible values for a 128-bit sequence than grains of sand on earth. Moving to 256 bits or 512 bits produces combinatorial sizes that are even less comprehensible. A sequence of 256 bits produces about 10^{77} possibilities; for 512 bits the set of possibilities becomes about 10^{154} .

Breaking encryption for these size sequences by trying all keys is impossible; there is not enough computational power to do so nor is it possible to create

enough computational power to do so. If keys or hashes are random then they are incredibly unlikely to be broken this way.

Attacking Encryption

It is impossible to try every possible key to attempt to decrypt something encrypted; there are too many keys to try as bounded by the laws of physics. For that reason, when users let their computers pick their keys or passwords randomly then decryption through automated guessing will not work. On the other hand, people are frequently not random when picking their own passwords, and they are often poor at securing the storage of those passwords. For example, passwords might be based on pets and birthdays and be written down in a notebook, or on a scrap of paper, or recoverable from a computer memory or disk. In these cases, there might be some hope that the password might be recovered. It is also possible to try and brute force passwords by repeated guessing. Humans often use statistically predictable patterns, and software exists that will work through trillions of combinations of statistically probable passwords to see if they create a key that can be used for decryption; this approach can be very effective for poorly chosen passwords, and very ineffective for randomly created ones.

In cases where the password is not recovered, it might be possible to get it from the user.²⁰ Some courts have reached a variety of conclusions whether a court can issue an order to compel disclosure of a password or even compel someone to unlock a device. Some courts hold there is no Fifth Amendment violation to compel a password's disclosure if the ownership of the device is a foregone conclusion.²¹ Other courts have explicitly refused to apply the foregone conclusion doctrine and denied orders to compel disclosure of passwords.²²

20. Orin Kerr, *Compelled Decryption and the Privilege Against Self-Incrimination*, 97 Tex. L. Rev. 767 (2019).

21. See, e.g., *United States v. Apple Mac Pro Comput.*, 949 F.3d 102 (3d Cir. 2020) (court upheld court order under the All Writs Act compelling disclosure of password under the foregone conclusion exception to the Fifth Amendment); *State v. Andrews*, 234 A.3d 1254 (N.J. 2020), *cert. denied*, 141 S. Ct. 2623 (2021) (the foregone conclusion exception allows the defendant to be compelled to communicate his memorized passcodes to the government); *Commonwealth v. Jones*, 481 Mass. 540 (Mass. 2019).

22. See *Seo v. State*, 148 N.E.3d 952 (Ind. 2020) (discussing concerns with extending the foregone conclusion exception to the Fifth Amendment in the context of compelling production of an unlocked smartphone); *Commonwealth v. Davis*, 220 A.3d 534 (Penn. 2019) (foregone conclusion exception to the Fifth Amendment is inapplicable to compel the disclosure of a defendant's password to assist the government to gaining access to a computer).

Cryptographic Hash Functions

A cryptographic hash is a mathematical function that takes any digital content as input of any length and produces a short unique numeric signature as output, typically 512 bits or fewer. This reduced number of bits, which is often referred to as the hash, is like a digital summary or identifier of the content that was input. Changing any single bit in the input changes the output hash significantly and unpredictably. We can therefore use hash functions to verify the integrity of large amounts of data, at any input size, at a level of granularity involving single bits. Good hash functions have the following properties that ensure the hash is secure:

- First, given a hash, it is not possible to recover the original data. This means that you can share the hash with anyone, and they cannot tell what was used to create it. If you give them the original data, though, it is easy to run those through the hash function to ensure that they are the same. Hashing is a one-way function.
- Second, given a hash, it is virtually impossible to find any other content that produces the same hash when run through the hash function. This property is related to the size of the hash output in bits: increasing the length of the hash output increases the difficulty of finding other content.
- Third, it is very difficult to create two different digital objects that have the same hash. This task isn't as hard as the second property, as we describe below, but it is still very hard for good hash functions. (By "hard" and "difficult" we mean that an adversary cannot acquire sufficient computing resources to achieve the desired result with even the smallest probability of success.)

Cryptographic hashes have myriad uses, the most common as digital summaries. To detect changes to a file, the hash of a file or other digital object is taken and recorded separately from the object itself. Later, the hash can be recomputed from the object or a copy of it. If the object has changed, then the hash will have changed. Similarly, no two objects will have the same hash value, as described above.

Accordingly, hash values are an important tool used in e-discovery as they provide a guarantee for the authenticity of an original data set and can be used as a digital equivalent of the Bates stamp used in paper document production.²³ Amendments to Federal Rule of Evidence 902 provide a procedure to authenticate electronic documents without calling the testimony of a witness. Committee notes to the 2017 amendment specifically point to the use of hash values to self-authenticate, relying only on a certification by a qualified person that they checked the hash value of the proffered item and that it was identical to the original.

23. Managing Discovery of Electronic Information 52 (Federal Judicial Center Pocket Guide, 3d ed. 2017).

Attacking Cryptographic Hashes

There are a variety of commonly used hash functions. Some older ones, particularly MD5 and SHA-1, have been shown to be weak or broken in ways that newer hashes such as SHA-256 are not. Specifically, attackers can violate the third property above and produce multiple data files that have the same cryptographic hash.

In contrast, for cryptographic hash algorithms that have no known weakness, the chance of generating a second file that has the same hash as a given, fixed first file is vanishingly small. For example, given a file with a particular cryptographic hash value that is n bits long and violating the second principle above, one would expect to have to generate 2^n candidate files before finding a match. For example, when using SHA-256, which produces a 256-bit output, one would have to generate 2^{256} (roughly 10^{76}) candidate files before expecting to find a match, which is essentially impossible in practice.

The process of creating hashes is subject to the so-called *birthday attack*, however, which violates the third property above. This attack is so named because while the chance of any one specific person having a set birthday is $1/365$ (ignoring leap years), the chances that some two people in a group of 23 having the same birthday is about 50%. In the first case, we have fixed on a single birthday date; in the second case, any matching birthday date achieves success. These are analogous to principles one and two above. Producing a large set of random digital objects will produce two objects that have the same hash more often than when fixing a file to match against. Specifically, one would expect to generate $2^{n/2}$ candidate files to find any pair that match from among all generated candidates. In the case of SHA-256, finding a matching pair requires generating 2^{128} (roughly 10^{38}) candidate files in expectation, which is still impractical.

Hash values are also used by electronic service providers to detect images of apparent child pornography. They typically configure their systems to watch for such images being transmitted or stored across their networks by computing hashes of the images and comparing them against hashes known to be illegal.²⁴ It is, however, easily possible to make alterations to the image that will not be visible but would change the hash.

Cryptocurrency

Cryptocurrencies have become the de facto standard form of conducting online payments for many illegal activities; in addition, there seems to be a growing

24. See *United States v. Miller*, 982 F.3d 412 (6th Cir. 2020) (analyzing Fourth Amendment concerns when Google used hash values to identify images of apparent child pornography transmitting across their network).

number of financial fraud cases involving cryptocurrency companies. In this subsection, we provide a very brief overview of how Bitcoin, the original cryptocurrency, works. There are many different cryptocurrencies that are based on a variety of cryptographic principles, but others are often similar in their design.

The fact that many cryptocurrencies have “coin” in their name is misleading. Instead of tracking individual currency units, most cryptocurrencies instead operate on the concept of a shared ledger. On this ledger there are a series of identifiers called addresses (analogous to a bank account number) that are each associated with an amount of cryptocurrency; the ledger is essentially a shared database that maps addresses to the cryptocurrency balance. This approach is as if a bank kept track of how much money is in an account without recording the serial numbers of any cash deposited and without attaching a name to the account.

Transactions cause some portion of a balance to be transferred from one address to another and recorded on the ledger. This ledger is called the blockchain and contains a history of every transaction that has ever happened in Bitcoin (or other cryptocurrency). By processing through the list of transactions listed on the blockchain, anyone can determine the balance associated with any address. In Bitcoin, each address is a public cryptographic key. Only the person who knows the corresponding private key may create transactions with that identity. Transactions are analogous to bank checks: they transfer currency controlled by one private key to the control of a different private key. One person might have control of any number of addresses. Transactions, however, are not anonymous in that they are inseparable from the addresses listed in the ledger. It is often possible to associate these Bitcoin addresses to real-world persons by finding Bitcoin transactions that are associated with real-world financial actions, like purchasing Bitcoin using a bank account or using Bitcoin to order physical items mailed to a postal address.

The major innovation of Bitcoin was to establish a mechanism by which a large group of computers could come to a consensus, without a central organizing authority, on which transactions become part of the blockchain despite some minority of participants attempting to cheat or otherwise influence the outcome. This mechanism, called mining, also creates new Bitcoin, and it incentivizes participation in the Bitcoin protocol, which helps make it resilient against attack.

The value people see in Bitcoin is that it is exceedingly difficult to alter the blockchain; to do so would require a very significant amount of computational power, in general roughly more than about half the computer power already performing mining on a chain; for the more popular cryptocurrencies it's most often an amount out of reach of any single entity. As long as a majority of participants, as rated by computational power, behave consistently then changing the blockchain should be quite challenging and therefore secure.

For some cryptocurrencies it is possible to anonymize the relationship between currency deposited in an address and later expenditures from that address. This process is often called mixing. Because the blockchain is a public ledger, it can be used by criminal investigators to follow illegal transactions. Blockchain

analytics can become key evidence in criminal activity involving virtual currency. For example, IRS investigators analyzed the blockchain and de-anonymized Bitcoin transactions allowing for identification of hackers compromising celebrities' Twitter accounts. The affidavit supporting the criminal complaint steps through the blockchain analysis leading to the identification of several co-conspirators.²⁵

Networking

The internet's core feature is the connections it creates among the world's people, computers, and devices. The internet is supported by an enormous amount of infrastructure, including: equipment commonly used in homes and offices, such as Wi-Fi access points and cable modems; equipment deployed by internet service providers (ISP) that have home and business customers; and the connections among ISPs that connect the world together. The internet is further extended by cellular network providers, including mobile phones, radio towers, and connecting infrastructure.

It has become commonplace for people to carry mobile phones and devices at all times, and for almost all communications to be based on the internet and modern cellular infrastructure. For this reason, a great deal of evidence and legal issues are based on network communications and services. To understand modern networks and the legal issues surrounding them, it's helpful to learn a number of fundamental concepts that we explain here.

Networking Layers

The connection of the world's cellular carriers, ISPs, and *end users* as one massive system is made possible by a collection of standards called the internet protocol (IP) suite. All computers that speak IP can communicate with each other regardless of the developer and manufacturer responsible for the software or hardware. These protocols are organized as several layers. At the top layer are the applications, such as email and the web. And at the bottom are the wires and cables that carry data. In between are a series of layers that handle reliability, routing, and access control, which are all terms that we define below.

An important aspect of this design is that each layer has its own naming and addressing scheme. An address is a place on the internet that data can go to, while a name is an identifier unique to that address. For example, an email "address" is actually composed of a name (like *alice*) connected by an @ symbol to an address (like *alice@umass.edu*).

To help understand the purpose and relationship among these layers, we can make an extended analogy to the postal service. Figure 2 illustrates the five IP

25. *United States v. Sheppard*, 3:20-mj-70996 (N.D. Cal. 2020), <https://perma.cc/5NSR-7QD3>.

layers. At the top is the application layer, which consists of protocols that operate between two instances of the same application running on different desktops, laptops, or smart phones. These different devices at the edge of the internet are often called end-hosts. Email clients, web browsers, messaging programs, social media applications, and online multiplayer games and other applications each send data across the network in their own way. Application-level content very rarely is examined or modified by devices on the network. One of the reasons the internet has been able to grow to a world-wide system is that what happens on these end-hosts is largely separate from the core infrastructure. In our postal service analogy, the application layer is like letters and magazines that are mailed: although the content is important to the sender and receiver, the actions taken by the postal system are not affected by the specific words, images, or other content within the letters, and the postal system generally does not examine the content.

Figure 2. The IP network architecture is divided into several distinct layers, each providing services that build on the layers below.

Application — email, web, gaming, instant messaging
Transport — ensuring reliable data delivery
Network — providing routing of data across the network and other services
Link — managing sharing of wireless and wired connections between computers
Physical — wires and wireless

Below is the transport layer, which also operates across the network between the two end hosts. The transport layer corrects problems and errors that are caused by the layers below so that applications see a consistent stream of data. The specific transport-layer protocol that performs these functions is called the Transport Control Protocol (TCP). In our analogy, TCP is like one of the special services that can be purchased from the post office, such as a delivery acknowledgment. For an important letter, you might make a copy of the letter and then send the original. If you don't receive a delivery acknowledgment after a time, you'll send a new copy of the letter and reset your timeout for the delivery acknowledgment; at its core, the algorithm providing reliable delivery of data operates the same way.

Next, going downward, is the network layer, which is the first layer that operates on the core devices that form the internet. Whenever information leaves an end host, it is handed off to a router. Just like desktops, routers are computers. The only difference is that they are not computers that are used to run just any program; they are specialized to operate programs that route data from one end host to another end host across a building or across the world. The entire internet is composed of routers that communicate in an unbroken mesh.

Most routers have more than one or two neighboring routers, and so just like roads that intersect with other roads, routers have to be navigated from a starting end host to a destination end host. A series of protocols determine for routers which of its neighbors most quickly leads to a destination. In our analogy, routers are like post offices. When a letter leaves a home, it travels from one post office to the next, until it reaches a final post office that delivers the mail to the destination home. The network layer is responsible for assigning an IP address, and it is often critical to attributing evidence in an investigation.

The link layer manages the connections between routers (and between hosts and routers). For example, the link between a host and its wireless access point can be managed by a Wi-Fi protocol. Stretching our analogy to almost its breaking point, links are like the postal workers that are the conduits for managing the many letters at a time that go from a home to the neighborhood post office, as well as transfers along a sequence of post offices on route to the destination.

Finally, at the bottom, physical layer protocols allow hosts to send data via some physical medium, including wireless, optical fiber, or coax wire. To complete our analogy, the vans, trucks, trains, and airplanes used by the post office are the physical objects that move mail from one point to another.

IP Network and Link Layer Addresses

The two most important computer addresses that are used to move a user's data across the internet are a computer's IP address and the address of a computer's network interface.

IP addresses are assigned by the local network administrator at an internet service provider (ISP). Before data can be transferred by a computer, it must obtain an IP address. They can be assigned once and remain static for the lifetime that the computer uses the ISP. Or, more typically, IP addresses are assigned dynamically by the internet service provider using the Dynamic Host Configuration Protocol (DHCP).

Network interface addresses are assigned at the factory that manufactured the radio card or Ethernet card, and they are typically 48 bits long. They are often written as colon-separated hexadecimal strings: e.g., 00:17:F2:40:F9:B2. The first 24 bits are a prefix that should identify the manufacturer; the second half should uniquely identify the hardware across the world. This value is easily

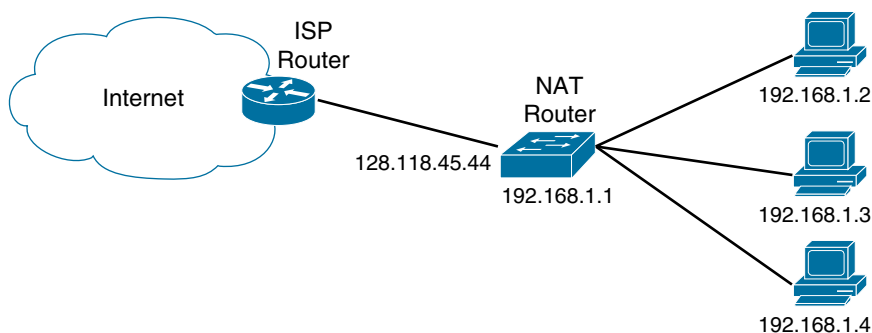
changed by the user, as we explain below. The link-layer protocols that make use of network interface addresses are called medium access control (MAC) protocols, and so most often the 48-bit values are called MAC addresses.

A person can take their laptop from one ISP to another. For example, a laptop may move from a person's home ISP to their work ISP, and then to an internet café afterwards, all in the space of one day. Each move will cause the computer to be assigned a new IP address—however, the MAC address will stay the same throughout. A computer's MAC address is never sent farther than a router or host one hop away, to the neighboring link-connected computers. Across the internet, it will be impossible to learn that one MAC address is behind all three IP addresses at different times.

Typically, a DHCP IP address assignment is logged, including the date and time, network interface address, and assigned IP address. The Communications Assistance for Law Enforcement Act (CALEA) asks ISPs to keep DHCP logs for 90 days, but this is not a requirement under the law. Some ISPs keep information for hours, some for longer than 90 days. Many law enforcement investigations involving a network are led by a search for the computer (and home) that has been assigned an IP address that was observed online. The ISP that assigned the computer needs to be found first, so that the DHCP logs can be requested via legal process.

Sometimes an IP address is not a completely unique identifier. There are only 3.7 billion of the shorter (version 4) IP addresses for the entire world, and they are assigned in large blocks by Internet Assigned Numbers Authority (IANA). Given the number of computers in the world, addresses are scarce. Some organizations have tens of computers yet are assigned only one address by their ISP. To get around this problem, many organizations (and home routers) make use of network address translation (NAT) boxes, which allow many computers inside a network to share a single external IP address. Figure 3 illustrates this operation.

Figure 3. Network address translation box hides a larger network behind it from the rest of the internet.

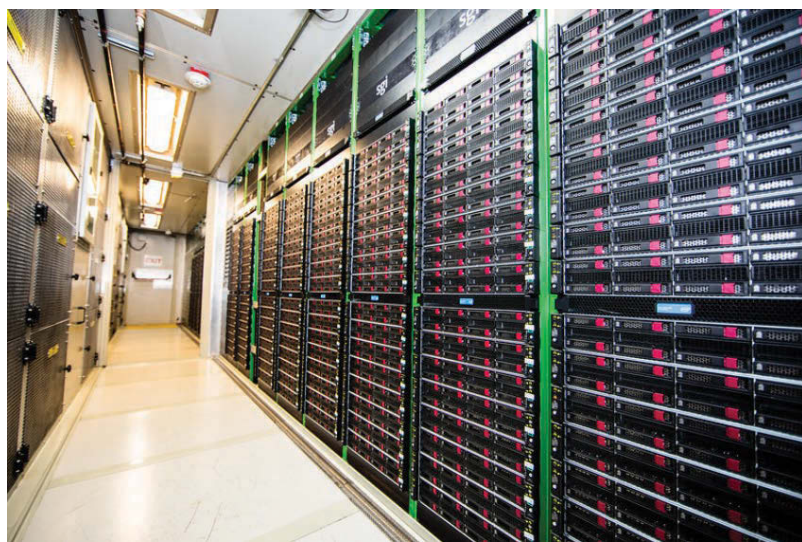


Cloud Computing and Serverless Systems

The term cloud computing refers to a popular approach to supporting online computer services. When the web and online services first started gaining popularity, it was common for businesses to operate a computer to serve online customers on their own premises. As the customer base grows, the business would not only require more powerful computing hardware, but investments in air conditioning, physical security, electrical power, redundant hardware, and more. Further, for some businesses, demand might be extraordinarily high for one day and not most of the year; it can be challenging to expand infrastructure for that one day.

Cloud computing addresses these issues for businesses. Cloud providers build large data centers that provide power, network access, and cooling for many thousands of computers. Typically, there are staff onsite to provide security and help deal with hardware problems. The computers used in these centers are different than laptops and desktops. Called rack servers, they are shaped more like elongated pizza boxes so they can stack neatly, as shown in Figure 4. They often include remote management interfaces so that they can be configured and operated across the network without a human needing to be present, but humans are required to address hardware problems.

Figure 4. Racks of computers or “rack” servers in a data center.



Source: Provided by the U.S. Department of Energy’s National Energy Technology Laboratory, <https://perma.cc/VGR4-URGD>.

When the notion of cloud computing was introduced, it was common to lease a complete virtual machine above, and this is still done. More recently, cloud providers realized that they could charge for hosting a specific program. For example, the program might accept a photographic image as input and return the text that appears in the image as output. This approach, called serverless computing, allows the cloud customer to avoid paying for an idle virtual machine. Instead, they pay for each execution of their program. The cloud provider ensures that as demand ebbs and flows, the execution of the program remains performant and reliable. Large providers of such services include Amazon Web Services, Google Cloud, and Microsoft Azure. Businesses can lease computing power from the provider and scale up or down quickly as demand for their product changes.

Cloud services are a good example of what is defined as a remote computing service (RCS) in federal law.²⁶

Wireless Networking and Mobile Devices

Wireless networking is commonly of two primary types. Cellular networks support long-range communication between a user's mobile device (e.g., a smart phone) and radio towers that can be miles away. Wi-Fi networks support short-range communication between a user's mobile device and access points (AP) that can be hundreds of feet away. Bluetooth networks are similar to Wi-Fi but typically operate over an even shorter range. There are many other types of wireless networks, including satellite-based communication, that we do not discuss here.

Wireless communications are typically broadcast in nature and based on radio waves. The physics of the radio frequencies involved and the expense of the equipment determines whether the signals can permeate walls or travel long distances. For example, cellular networks are based on very large antennas situated on very large towers. Wi-Fi networks are based on small antennas affixed to boxes placed on shelves in homes. The difference explains the large geographical ranges that cellular networks can traverse compared to Wi-Fi networks. Given that wireless signals will be received by anyone within range, cryptography is typically used to keep communications confidential between the two parties communicating over the wireless channel.

26. See the definition of remote computing service, 18 U.S.C. § 2711(2).

Cellular Networks

Cellular networks are deployed across large geographic areas to allow for a user to be mobile and stay connected. Originally, cellular networks were deployed to support voice phone calls. Over time, the networks have evolved to support the data connections required by internet-enabled apps and services common on smart phones.

To enable cellular service, towers or other installations with antennas are connected to the provider's network over backhaul connections that go over cable, fiber optic, or microwave links. Each tower that a cellular provider deploys covers a geographic cell in which users have connectivity. The towers are often called cell sites. The towers are connected to a complex set of systems that route data to users and ensure that only paying, authenticated customers have access. Commonly, a subscriber identity module (SIM) card is issued by the cellular provider and inserted into the user's mobile device. The SIM card uses cryptography to authenticate the user's subscription. Records held by the cellular provider link the user's identity and billing information to the SIM card. The mobile device has an International Mobile Equipment Identity (IMEI) assigned to it.

As cellular networks have advanced in terms of range and available bandwidth, they have changed names, including 2G, 3G, LTE, and 5G. These names represent a combination of changing technical specifications and marketing efforts. The basics have remained the same even if internally the industry has changed its terminology continually. Therefore, we will describe these basics in high-level terms.

When a user powers on a mobile device, the device seeks out a nearby cell site. Using the credentials held in the SIM card and via a well-defined signaling protocol, the device authenticates to the cellular provider. As the user moves, the device will associate to a new tower with a stronger wireless signal. Voice calls and data are routed to the tower the device is associated with. Typically, the provider stores a record of these associations called cell site location information (CSLI). If a device is in range of three or more towers, the device's geographic position can be estimated from the signal strengths. This estimation can be performed after the fact if the signal strengths are recorded, or it can be performed in the moment and the estimation stored in a record. Alternatively, if the mobile device contains Global Positioning System (GPS) functionality, it can tell the cellular provider of its location. Geographic location is not required to provide wireless communications; it is required to assist with calls to emergency 911 services.

The internet includes mobile devices attached to cellular networks. These devices typically are assigned an IP address via the cellular provider's NAT gateway. Some cellular providers keep records of these NAT assignments; some do not. The amount of time these records are kept by providers also varies.

Wi-Fi

Wi-Fi is a commercial term for a series of industry standards for short-range wireless communication among computers and mobile devices. Wi-Fi networks are inexpensive for consumers to deploy to cover part of a living area with one central access point (AP). With several coordinated APs, a large home can be covered. The same technology can be scaled up to cover a large building or campus. Typically, the Wi-Fi AP includes NAT and security functionality.

Many consumer devices are available that serve a single purpose and are designed to connect to the user's Wi-Fi network. These so-called internet of things (IOT) devices include services such as monitoring and adjusting temperature and lighting, controlling door locks and entryway cameras, and managing smoke alarms and other sensors. Some IOT devices connect to each other, smartphones, and desktop computers via Bluetooth wireless technology. Bluetooth was designed to allow for very short-range connections between devices, replacing short cables. Bluetooth operates very differently than Wi-Fi internally, but it includes the same concepts, such as a MAC address and broadcast radio waves.

Network Architectures

The classic architecture used for coordinating the communication of a large set of computers over a network is called client-server. A single server acts as the single point of coordination of all communication, and the multiple clients never transmit data to one another. The burden of work is on the server, which must be a machine that is provisioned with significant network bandwidth, disk storage, and processing. However, the reliance on a single point of coordination results in a system that is easier to control and manage, i.e., there is little complexity to the clients' operation and design. Figure 5 illustrates this architecture.

Peer-to-peer (p2p) network systems are possible because home computers have sufficient resources, having access to high-bandwidth home internet connections. Peer-to-peer systems allow a group of clients to pool their resources and offer a networked service to each other, even though no single client has sufficient resources to act as a server. The disadvantage of peer-to-peer is an increase in complexity at each peer: each plays the role of a client and a server. The details of how a particular peer-to-peer network operates can be complex for that reason. Figure 6 shows a possible p2p architecture.

Peer-to-peer applications are a popular way of sharing content with other users without relying strongly on a centralized server. Because there often appears to be no central computer in charge, many users of p2p applications believe that they have relative anonymity. While it is more difficult for investigators to examine these systems, they are not anonymous, as the efforts of the recording

Figure 5. A client-server network architecture. Dotted lines represent connections across the internet.

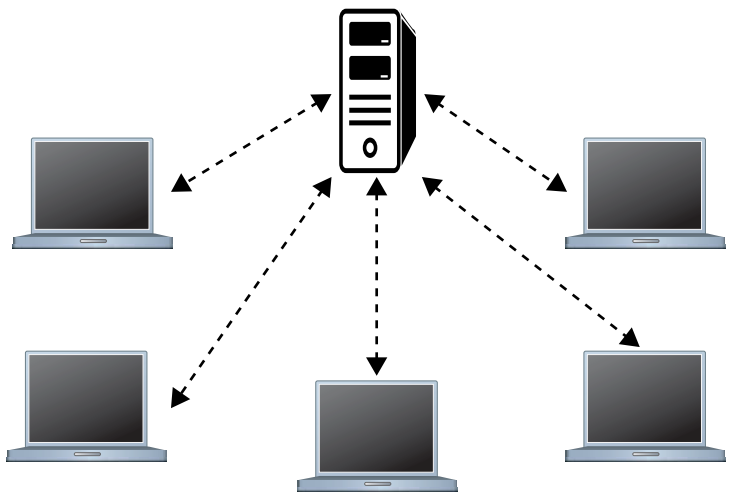
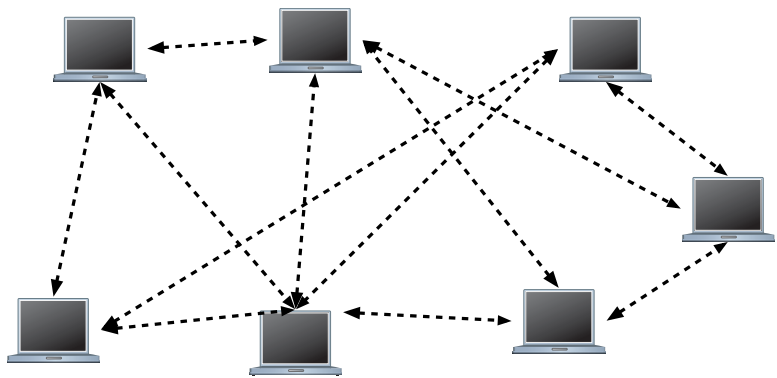


Figure 6. A peer-to-peer network architecture. Dotted lines represent connections across the internet.



industry have shown.²⁷ We discuss applications specifically designed to provide anonymity on the internet, including Tor, below.

Peer-to-peer applications are widely instrumental for the sharing and distribution of files containing copyrighted music, videos, and software, as well as child sexual abuse material (CSAM). BitTorrent is the primary p2p file-sharing

27. A&M Records, Inc. v. Napster, Inc., 239 F.3d 1004 (9th Cir. 2001).

application in use on the internet today. Others include eMule (also called eDonkey), Ares, and Gnutella.

Anonymity and Obfuscation

Whenever a computer uses the internet to directly connect to another computer, its IP address is revealed. To obfuscate their address, one of the senders could use a third party. For example, Alice may communicate to Carol with the help of her friend Bob as an intermediary. In this subsection, we explain the most prominent methods used for obfuscating IP address information.

Open-Access Wi-Fi

The simplest method of obfuscating one's real IP address is to make use of open-access internet connections, such as free Wi-Fi offered by cafés and other businesses, or by libraries and other municipal services. Free Wi-Fi is almost always provided using network address translation (NAT); see section titled “Network Address Translation,” above. An inexpensive Wi-Fi access point providing the connection and the NAT service that might be used by a small business or home typically do not have logs of the device that was provided service. For someone seeking obfuscation from free Wi-Fi, the disadvantage compared to other mechanisms is that they must be in physical proximity of the Wi-Fi. Their image or their vehicle's image may have been caught on security cameras. And some Wi-Fi services require registration via an email address, which can be a lead to start an investigation of illegal activity.

Virtual Private Networking (VPN)

Virtual Private Networking (VPN) services work similarly to hide IP addresses. In this example, we describe how entities can communicate while hiding their IP address. We do this in a manner common to computer science, where entities referred to as “Alice” and “Bob” and others participate in communications.

Alice can hide her IP address from Carol's computer if she sends her communications through Bob's VPN service. Alice's computer connects to the VPN endpoint (also called a VPN concentrator) administered by Bob, and all traffic she sends is encrypted so that it cannot be read or altered by others. Many businesses make use of VPN connections to allow their employees to work remotely. In these scenarios, “Bob” is the company that “Alice” works for. A VPN connection from an employee's home computer to systems administered by the business ensures the communications over the internet are confidential. Furthermore,

such connections can involve credentials that are issued only to the employees, ensuring the connections are authorized.

In other scenarios, “Bob” is a company that sells VPN services as a proxy that offers privacy to users like “Alice” who pay for the service. When Bob’s VPN endpoint receives Alice’s communications, he sends it to Carol’s computer with his own IP address as the sender. Thus, Carol learns the IP address of Bob and not Alice. In these scenarios, Alice trusts Bob with the knowledge that she is communicating with Carol. While Alice has an active communication with Carol, Bob’s VPN endpoint has to keep track of the connection so that return traffic from Carol can be routed to Alice. Once the communication is over—for example, after a single web page is retrieved—the record of the connection can be disposed of. Many commercial VPN service providers advertise a “no logs policy,” which means that the record of past communications is not kept by the VPN service provider. Bob is providing a more secure service for Alice by erasing logs that might be discovered by third parties that seek to monitor Alice’s communications. If Alice’s interest in Bob’s services is to obfuscate her illegal activities, then a no logs policy is similarly advantageous. For example, VPN services are known to be hurdles for investigations into child exploitation.

Anonymous Communication Systems

Several services on the internet were designed to provide stronger obfuscation than free Wi-Fi and VPN services. The Tor Project has created software that is used by volunteers on thousands of computers worldwide. Each computer is called a relay. This peer-to-peer network of relays is called the Tor Network and supports two systems for anonymous communication. The first is Tor Browser,²⁸ which is by far the most prominent system that can be used to browse internet websites without revealing the client’s IP address. The second is Tor onion services, and it is used to prevent disclosure of a server’s IP address. A variation of the approach used by the Tor Project is employed by the Invisible Internet Project (i2p), though it is less popular than the Tor Project. The i2p system comprises a set of relay computers that are distinct from the Tor Project. Freenet is another lesser-used obfuscation service that operates differently from the Tor and i2p systems.

The term “darknet” or “dark web” is often used by popular media to describe services that obfuscate a user’s IP address, but it is a poor word choice.²⁹ The term

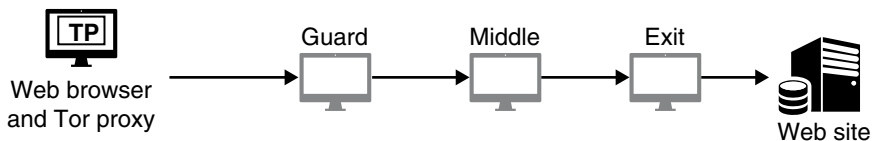
28. Roger Dingledine, Nick Mathewson, & Paul Syverson. *Tor: The second-generation onion router*, in Proc. Conf. on USENIX Security Symposium, 2004.

29. This text based in part on “Increasing the Efficacy of Investigations of Online Child Sexual Exploitation: Report to Congress,” Brian Levine, May 2022. NCJ Number 301590, <https://perma.cc/WA8J-SURQ>.

darknet originally referred to the fact that the content made available by these services is not indexed and made searchable by sites such as Google and Bing, hence the content is “dark” and not easily found. It is more instructive to consider the purpose and structure of the services. The Tor Network is designed to thwart the ability to attribute the content of traffic from the user. VPNs, which we describe above, are single-proxy systems for obfuscation. Tor and Freenet are multi-proxy anonymous systems for IP address obfuscation (anonymous systems, for short). Tor and Freenet are peer-to-peer networks, and they are possible because volunteers operate internet-connected computers running the software. Software for Tor and Freenet is free and does not require users to register or identify themselves with a central authority. Consider that in contrast to using a VPN where users typically register with a VPN service provider, likely submitting a form of payment such as a credit card number.

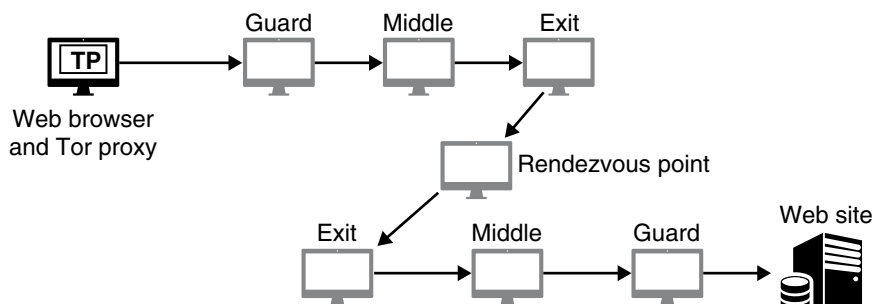
Tor Browser is similar in function to a single-proxy VPN service in that a remote web server can be contacted without revealing the IP address of the Tor user’s computer, as illustrated in Figure 7. Unlike a VPN service, Tor Browser uses three Tor relays in sequence. Communication is encrypted in layers so that each relay knows only the previous and next steps in the chain. The clear internet destination of the communication is known to only the last relay, called an exit node while the exit node does not know the IP address associated with the initial sender. The destination web server is unable to learn the IP address of the user. The guard node does not know the destination’s IP address; the exit node does not know the user’s IP address; the guard and exit are separated by the middle node. Tor relays do not keep logs about the connections formed, and because the service is operated by volunteers it is free. Tor Browser provides anonymous internet connectivity in that it hides the IP address of the user’s computer but does not modify the content sent.

Figure 7. Tor Browser connects a user to a website in a multi-proxy setting.



Tor onion services, previously called “hidden services,” allow for a website (or any server) to hide its IP address from users that communicate with it. The site is behind three relays, awaiting web requests. As illustrated in Figure 8, the requests can come from Tor Browser users only. (A version of onion services is present in i2p, and they are referred to as eepsites.)

Figure 8. Tor onion services connect a user to a website in a multi-proxy setting. Unlike Tor Browser, the IP address of the website is unknown to the user.



Freenet³⁰ is much less popular than Tor, but it is used daily by thousands of users. Freenet provides an anonymous service where users can publish and retrieve files. It cannot be used like Tor Browser or a VPN service to connect to the clear internet. Freenet can only retrieve content that has been previously inserted in Freenet's system by other users, and in that sense, it is akin to Tor onion services. Freenet attempts to prevent attribution of downloads using an approach that is different from Tor; we do not detail its operation here.

Network Identifiers and Third-Party Data Collection

All told, a tremendous number of layered mechanisms allow for the internet, mobile communications, and modern apps and websites to work seamlessly and reliably. Each layer typically makes use of various network names and addresses that uniquely identify a device or its user. The IP address of a device is the most obvious example, but many more are present and often contribute to the evidence in a case. Below we provide a glossary of the more common identifiers.

IP addresses used on the public internet are attributable to a particular internet service provider (ISP) via public records. ISPs are given a contiguous range of addresses in a block, which they can assign to their customers. As described above, it's possible that one address can be shared among many users as part of a NAT setup, whether within a home or business. Often IP addresses are associated

30. Ian Clarke et al., *Freenet: A distributed anonymous information storage and retrieval system*, in *Designing Privacy Enhancing Technologies*, pp. 46–66, 2001, https://doi.org/10.1007/3-540-44702-4_4.

with a billing record that has a street address and credit card information. Logs of the assignment of an address to a customer usually exist, but no law requires that they be stored or retained. For example, many VPN providers do not keep records.

Email addresses are one of the oldest identifiers on the internet. Typically, email addresses are secured by an email provider with a password, and logs of access to the account often exist. It is a challenge to authenticate the true sender of a received email from the email alone. A log of the sending of the email or a copy on the sender's computer is stronger evidence.

Cellular phones have a series of identifiers. These include the following: unique phone numbers, which are assigned by the provider; International Mobile Equipment Identity (IMEI) that uniquely tag a hardware device; and the International Mobile Subscriber Identity (IMSI) that uniquely tags a particular SIM, which is a small card containing tamper-proof electronics issued by the cellular provider and inserted in the mobile device. Typically, providers keep track of all three values (and more) and the calls and connections to cell tower base stations for each device.

Almost all computers are connected to the internet via a shared network at the link layer (see section titled "Networking Layers," above). Wi-Fi and cellular connections are almost always shared with other computers. Wired Ethernet connections are also typically shared. In all these cases, each device on the shared wireless or wired medium is identified by a unique medium access control (MAC) address. While usually globally unique, a MAC address can be changed quite easily by the user and is often not shared beyond the local network.

As users browse the web, the websites they visit store small amounts of information within their browsers. Generally, this information is called a cookie, and more broadly there are many ways for a website to store data on a user's device. The stored information can be useful for the website to reidentify visitors and store preferences. Some mobile phones and some desktop devices are assigned unique advertising IDs (AdIDs) and other identifiers that identify the user's devices across different apps. These values can be cleared easily by the user, but many users do not.³¹

Within some systems, there may be other identifiers that are stored internally to track user activity. For example, web-based systems that require users to log in typically assign each user an internal identifier. The identifier is often a number that is used to look up the user's information in a database. Each action the user takes, which can vary by system, is linked to this identifier. For example, posts in social media systems will be linked to this identifier. The identifier is typically logged each time users perform an activity, from creating the account

31. Keen Sung et al., *Re-identification of mobile devices using real-time bidding advertising networks*, in Proc. ACM International Conference on Mobile Computing and Networking, September 2020, <https://doi.org/10.1145/3372224.3419205>.

to each time they log in, and each post or comment they make. Given some information about the user, a site administrator should be able to retrieve the business information they have collected on the user. The resulting data will vary by what each site has determined to collect and keep.

Most of the identifiers above are assigned or stored by third parties (e.g., advertisers and app platforms), as their purpose is to globally distinguish the user or their device. Thus, it can often be key information in both civil and criminal cases. It is common for these values to be both stored on a device and with a third party. The legal process required to obtain the information of course depends on many factors, including who holds the information sought, whether the information is provided in real time during transit or if it is stored prior to disclosure, and how the information is categorized by law (e.g., content, subscriber information, or other records). Several statutes may come into play here, including the Stored Records Act (18 U.S.C. §§ 2701–2710); Pen Register and Trap & Trace statute (18 U.S.C. §§ 3121–3126); and the federal wiretap act known as Title III (18 U.S.C. §§ 2510–2522). The Department of Justice’s Guide to Searching and Seizing Computers³² is a good resource to step through the application of these laws to obtain the various forms of identifying information discussed here.

Geographic information is a part of many networking systems and apps at various levels of granularity. Global Positioning System (GPS) information is the most accurate type of information. GPS coordinates are determined by devices based on signals received from orbiting satellites. The accuracy of a position can be within tens of feet under perfect conditions. However, tall buildings and other occlusions can reduce the accuracy.

Not to be confused with GPS is the concept of IP address geolocation, which is a method by which the geographic location of a device is estimated by its IP address. Geolocation accuracy varies widely depending on many factors. It can be accurate for IP addresses assigned for a long duration of time to places with a fixed location, for example the IP address assigned by an ISP to a home. IP address geolocation can be inaccurate and misleading if used incorrectly,³³ and it is inherently inaccurate for devices on cellular networks or connected to a VPN service.

The geographic location of a mobile device might be discovered via cell site records kept by cellular providers. As devices roam the cellular network, they attach to radio towers (also called cell sites) that have the strongest signals. Since the towers are in fixed locations, knowing the tower a device was attached to provides information about the device’s location. If several towers are within range of a device, then it is possible to estimate the location more accurately. Specifically, a device’s location can be triangulated among three or more towers. To do

32. Available at <https://www.justice.gov/criminal-ccips/ccips-documents-and-reports>.

33. Cyrus Farivar, Kansas couple sues IP mapping firm for turning their life into a “digital hell,” *Ars Technica*, Aug. 10, 2016, <https://perma.cc/MUZ4-YMFM>.

so requires the signal strength information from the towers, either in real time or as part of a stored log.

It can be instructive to consider the context of each of these identifiers and other information that is a part of typical computer use and internet communications. Dialing, routing, addressing, and signaling information is abundant in network communications. When real-time dialing, routing, addressing, and signaling information is collected, the Pen Register and Trap and Trace Statute may be implicated. *See* 18 U.S.C. §§ 3121–3126. While IP addresses may be considered routing information, many identifiers addressed here are not.

Putting It All Together

A short example can illuminate how many of the technologies detailed above fit together to leave a trail of evidence about a user's actions.

Alice begins her day by checking her mobile phone. She unlocks her phone with a biometric (such as her face scan or fingerprint). She realizes that her phone is disconnected from her home's Wi-Fi and her cellular provider. She connects to her home Wi-Fi (but not her cellular provider) and immediately several apps are active. Her email client retrieves new email messages sent overnight. Other apps send and receive notifications related to social media messages, sales, and promotions. A device worn on her wrist has already synchronized information about her sleep duration with her mobile phone via Bluetooth. From these actions, several companies around the world may have stored records that associate her user accounts with activity from her home's IP address at a specific time. Notably, Alice's home ISP has not changed her IP address in many months.

Before she leaves home, she browses social media and reads the news. Advertisements make note of the device's AdID and IP address. She has allowed her weather app access to her geographical location and both her location and AdID are relayed to hundreds of third parties who bid to show her an advertisement.

As she leaves home to go to the store, she realizes that she needs to connect to her cellular provider and turns on the cellular connection, possibly by turning off airplane mode, which disables the cellular connection for flights. A record now exists of her device's association with a specific cell site. As she drives to work, the series of cell sites that her mobile device connects to is recorded by her provider. Her navigation app shows her more advertisements; the AdID is the same as when she was home even though her IP address has changed.

She enters a store, and her phone connects to the Wi-Fi. She purchases some items on credit cards stored on her phone and Bluetooth.

During this short excursion, she was captured in videos taken by her own front door camera and the store's cameras. A neighbor down the street has a camera that captures her car drive by. Her financial transactions are stored on her phone, at the store, and with the credit card company. She has sent text messages

on several apps providing some context to her trip; these messages are on her phone, stored in the cloud, and on friends' phones. A fitness device captures aspects of her movements.

Because she shares access to her mobile device with no one, there is reason to believe the messages are truly authored by her, and that other aspects of the records are accurate. The fact that the records are stored by many third parties also speaks to their accuracy and integrity.

Investigating Network Traffic

Investigators can acquire evidence relevant to civil or criminal litigation by examination of network traffic. In some cases, investigators can, with permission from a court, directly monitor the traffic being sent over a wire or a radio link and capture a recording. In others, they participate in ongoing network communications and, as participants, receive the network traffic shared among all participants.

Network engineers and administrators sometimes need to observe network traffic to help understand and debug network issues. For this, they use one of many tools that can read, record, and interpret all traffic coming across some communications link. One such tool is named Wireshark, which is freely available and commonly used. Wireshark can collect packets traversing a link, record them in a commonly used format known as a packet capture (or pcap) file, then analyze the traffic contained in the file. This collection provides a complete record of what was sent and received, including the participant addresses and data within the packets.

Investigators gathering information on network communications can use the same tools and methods to record and analyze network traffic. Given access to a network link, investigators can record the network traffic and then analyze it to determine what communications took place during the observation period. This can reveal the addresses of the parties that were using the link for communication and the data that were being sent between them. Such collection may be lawfully done when an investigator is a party to the communication. In such cases the investigator can record all network communications sent to their device, as they are one of the intended recipients. This is known as consensual monitoring.³⁴ Another relevant exception to the federal wiretap statute that authorizes the collection of network traffic without a wiretap order is known as the computer trespasser exception, which authorizes law enforcement to collect the communications of a computer trespasser transmitted to, through, or from the protected computer if the owner or operator of the protected computer authorizes the collection.³⁵

34. See 18 U.S.C. § 2511(2)(c) & (d).

35. *Id.* § 2511(2)(i).

Investigating Peer-to-Peer File-Sharing Systems

As described above, investigators can gather evidence from public activity conducted on peer-to-peer networks. Different p2p networks have different utility to users; some, for example, exist to share files between users, others to facilitate cryptocurrency operations. In general, anyone can choose to join a p2p network, and the network makes use of their computer's computational power and network bandwidth and includes it in the operation of the network. As part of the network, each computer processes data voluntarily sent by other computers in the network. Often, the data can include evidence of illegal activity.

For example, p2p file sharing networks allow users to exchange copies of files; in some cases, they support searching for particular files as well. Well-known p2p file sharing networks include BitTorrent and Gnutella. Commonly, the files being shared are subject to copyright. In other cases, the files being shared can contain child sexual abuse material.

Civil or criminal investigations on these networks are aided by the fact that normal operation of the p2p network results in peers sharing information voluntarily with other peers. Any investigator who connects to a p2p network as a user will thus be able to see some of the activity that is happening on the network. For example, peers often publicly advertise the files they are sharing or seeking to obtain. Often files are identified on file sharing networks by the hash value associated with the file. The file sharing network is designed to allow a user to then download the shared file. When an investigator has joined the file sharing network as a user, they can then download that file directly from a particular peer to confirm possession and sharing of that file. As part of that download, the investigator's computer will see the network address information, which can be used to later identify the user.

All this activity occurs publicly as part of the normal operation of the network. No special access is required by the investigator. Courts considering civil or criminal cases involving p2p networks have consistently held that the information shared on the p2p network is publicly available, including but not limited to the user's IP address and shared files.³⁶

Investigating Anonymous Communication Systems

Anonymous systems were designed with the intention of providing privacy for persons who are vulnerable and require protection, such as journalists. However, for decades, anonymous systems have been heavily leveraged by those who

36. See, e.g., *United States v. Borowy*, 595 F.3d 1045 (9th Cir. 2010).

commit crimes such as child sexual exploitation,³⁷ creating malware botnets,³⁸ and selling illegal drugs.³⁹ Investigations of anonymous systems make use of a variety of techniques that are tailored to the operational details of each system.

Users associated with criminal Tor onion services have been identified by law enforcement when the criminal sites and information contained on those sites were seized by law enforcement. For example, in 2015, the FBI obtained a search warrant to install computer code on the seized child exploitation onion service named “Playpen.” When users visited the site, the computer code inserted and executed code on the visitor’s computer, which then conducted a remote search of the user’s computer then sent the true IP address of the user to the FBI (see section titled “Legally Authorized Intrusions,” below).⁴⁰ In 2017, the Dutch National Police obtained judicial authority to seize and take control of the “Hansa marketplace,” an onion service operating as a marketplace for illicit goods and services including drugs, firearms, and cybercrime malware. While operating the online marketplace, law enforcement was authorized to monitor the criminal activity occurring there, which led to the collection of information on high-value targets and delivery addresses for tens of thousands of orders.

Investigations of users of the Freenet anonymous network are based on the network traffic that travels between neighboring Freenet peers.⁴¹ As described in the section titled “Network Architectures,” above, Network Architectures, users of the Freenet system can retrieve content that has been previously uploaded by others into the network. Users direct their Freenet software to request a desired file from neighboring peers, one piece at a time. Neighbors of the original requesters either provide the requested file piece, or they relay the request to one of their neighboring peers. The relayed request contains information about

37. Elie Bursztein et al., *Rethinking the Detection of Child Sexual Abuse Imagery on the Internet*, in The World Wide Web Conference 2601–07 (May 2019), <https://doi.org/10.1145/3308558.3313482>; Rebecca Sorla Portnoff, *The Dark Net: De-Anonymization, Classification and Analysis* (March 2018) (Ph.D. dissertation, U. Cal. Berkeley); Clement Guitton, *A review of the available content on Tor hidden services: The case against further development*, 29 Computers in Human Behavior 2805–15 (Nov. 2013), <https://doi.org/10.1016/j.chb.2013.07.031>; U.S. Dept. of Justice, *The National Strategy for Child Exploitation Prevention and Interdiction: A Report to Congress* (April 2016); Brian Neil Levine, *Report to Congress: Increasing the Efficacy of Investigations of Online Child Sexual Exploitation* (National Institute of Justice) (May 2022), <https://www.ojp.gov/library/publications/increasing-efficacy-investigations-online-child-sexual-exploitation-report>.

38. Gareth Owen & Nick Savage, *Empirical analysis of Tor Hidden Services*, 10(3) IET Information Security 113–18 (May 2016).

39. Press Release, U.S. Dept. of Justice, AlphaBay, the Largest Online “Dark Market,” Shut Down (July 20, 2017), <https://perma.cc/8KFQ-R3PS>.

40. See Press Release, U.S. Dept. of Justice, Florida Man Sentenced to Prison for Engaging in a Child Exploitation Enterprise (May 1, 2017), <https://perma.cc/82U9-NQRU>.

41. Brian N. Levine, et al., *A forensically sound method of identifying downloaders and uploaders in freenet*, in Proc. ACM Computer and Communications Security (Nov. 2020), <https://doi.org/10.1145/3372297.3417876>.

what is being requested along with other distinct signaling information necessary to operate the network as designed. Careful statistical analysis can be done to distinguish whether a neighboring peer requesting content is the peer that made the original request or merely a relayer.⁴² When law enforcement joins the Freenet network, it receives requests for illicit material just as any other user of the Freenet network would. (Notably, as the investigator is the intended recipient of all information used in the analysis, the investigation is not similar to the techniques and legal authorizations relevant to the Playpen cases described above.) Using statistical analysis, law enforcement can then calculate the probability the neighboring peer sending the request for illicit material is the original requester and can move forward with an investigation.

Computer and Network Security

In this section, we provide an overview of computer and network security principles to help judges understand the nature and extent of cybercrime, including hacking, malware, and other types of cyberattacks. We provide an overview of privacy issues to aid judges in determining whether data breaches are the result of insufficient security measures and, if so, whether they constitute a violation of data privacy laws. These principles might also arise in intellectual property cases when considering the theft of trade secrets or other work, or in national security matters to understand how attacks on critical infrastructure can threaten the nation.

The classic definition of security has three aspects: confidentiality, integrity, and availability. Confidentiality refers to preventing access to information by those who are not allowed to see it or know it. Integrity refers to the reliability of data. In some cases, this refers to ensuring that the accuracy of data is maintained; it can also refer to digital data not being altered by those who are not allowed to do so. Availability means that we can access systems and data when required and that data cannot be destroyed by those not allowed to do so.

Each of these fundamental aspects relates to some policy that defines who is allowed to perform some action, like accessing, changing, or deleting data. An important aspect of digital security is therefore being able to prove who you are so that access to the appropriate actions can be granted. This is called authentication. Additionally, the idea of nonrepudiation means that users cannot deny their actions and that they can be held accountable for them.

It is worthwhile to note that the desired security aspects depend on the situation. For example, the government would like classified intelligence documents to remain confidential, unaltered, and available. Confidentiality is not required, though, for public governmental websites, though it would be desirable for them

42. *Id.*

to remain accurate and available. Some instant messaging programs, notably Snapchat and Signal, have messages that are automatically deleted after some period and thereby gain confidentiality at the cost of integrity and availability.

It is common to approach assurance as a cycle with at least three steps (though more detailed models exist): systems are *protected*; mechanisms then attempt to *detect* when protections have failed; and then people *respond* to the event. The response can include upgrading protections, thus completing the cycle. This cycle is common in many security situations. For example, we protect our homes with locks and fences; detect when those have been bypassed by cameras, barking dogs, or alarm systems; and then recover from incidents by summoning the police, relying on insurance, or getting better locks.

It is important to be aware that detection, like protection, is an imperfect process. Most people are familiar with this in the context of car alarm systems. Almost no one responds to car alarms because most often it is a false alarm; the alarm sounds even though no one is breaking in. This is called a false positive. Similarly, sometimes cars get stolen without the alarm going off, which is called a false negative. All detection systems have some rate of false positive and false negatives.

In addition, detection systems suffer from the base rate fallacy, which occurs when the thing a system is trying to detect is very rare compared to the number of events. An example might be a system to differentiate terrorists from normal passengers on airline flights. Assume there might be 1,000 terrorists who fly every year, and we have a system that can detect them with 99% accuracy. This system can also identify normal travelers with 99.9% accuracy, and in the United States there are about 400 million travelers a year. (We note that these are exceptionally accurate systems; it would be quite difficult to reach these numbers in practice.) This system would correctly identify 990 travelling terrorists a year ($1000 * 0.99$), but it would also falsely identify 400,000 normal passengers as terrorists ($400,000,000 * (1 - 0.999)$). The problem is that only about 0.25% ($990 / (400,000 + 990)$) of the resulting alerts are correct; the sheer number of misidentifications of normal travelers dominates the alerts.

The resulting low positive predictive rate, which represents what part of alarms are useful, might not be a problem depending on the circumstance; it might be this initial screen is fast and inexpensive and that a more costly test is used on the positive results. A risk, however, is that humans who monitor the alerts will become habituated to the large number of false alarms and miss correct ones.

Privacy

Privacy online differs from security because privacy is concerned with what information about you others might possess and share; security is primarily concerned with protection of your own data. The advent of online discourse and commerce has allowed a variety of entities to collect, store, and share information about users

that visit their sites. Advertisers and others are able to trace an individual across multiple sites and build a profile of that user and their preferences and interests. They can then use this profile for many purposes, including trying to identify items to promote for sale, or influencing political views and activities. There are many technical ways to identify users across sites; even sophisticated users who try not to be tracked can fail.

Online privacy is addressed in numerous laws and regulations; however, individuals often agree to the disclosure of their data to others by agreeing to the terms of use before using software, websites, or apps. Class actions have arisen when companies are accused of disclosing data in violation of their terms of use agreed upon by the users. For example, in 2022, Facebook's parent company, Meta, agreed to pay \$725 million to settle a class action privacy lawsuit where plaintiffs alleged in part that Facebook shared user data with business partners without disclosing such sharing and failed to restrict and monitor third parties' use of Facebook users' sensitive information.⁴³ In addition, a company's computer systems may be hacked, resulting in the release of private information that users never expected nor agreed to have released.

Sometimes poor cybersecurity can lead to violations of privacy laws. For example, in 2017, Equifax experienced a breach that resulted in the loss of personal and financial information for nearly 150 million people owing to its failure to fix a vulnerability in software associated with one of its databases. The Federal Trade Commission (FTC) filed suit against Equifax alleging violations of the FTC Act, 15 U.S.C. § 45, and the Safeguards Rule codified at 16 C.F.R. Part 314 issued pursuant to the Gramm-Leach-Bliley Act (GLB Act) by failing to reasonably secure the sensitive consumer personal information held in their computer networks. The parties reached an agreement resulting in a stipulated order for a permanent injunction and monetary judgment. The injunction prevented Equifax from making misrepresentations regarding the protections of personal information, ordered the creation of an information security program and third-party assessments of the information security program, and ordered prompt reporting of any future unauthorized access to a consumer's information. The monetary judgment totaled \$575 million, with the majority of this judgment used to provide direct relief to consumers by compensating for related harms and providing extensive free credit monitoring.

Authentication

An important part of computer security is accurately identifying a user so that the correct security policy can be applied to their actions. In particular, evidence

43. Facebook, Inc. Consumer Privacy User Profile Litig, 3:18-md-02843-V (N.D. Cal.).

of user actions might be stronger or weaker depending on what form of authentication was used. A password or device might be stolen and reused; biometric data might show a person was present.

What You Know (Passwords)

The most common and least expensive form of authentication is something you know, in the form of a password that is shared between the user and system. Passwords have many vulnerabilities, however. Anyone who knows a password can use it. An attacker can use an automated process to try likely passwords to gain access to accounts—and in fact do so commonly. Users also often reuse passwords across accounts. If one account is compromised and the password stolen, the same password can be tried at other sites to see if it works there. Such password reuse attacks are also common.

Some organizations use a password recovery process that has the user answer questions about themselves when the account is created. These questions are often used as an alternative to a forgotten password. The information used in questions is sometimes guessable, and an attacker with time to research their subject can often answer them. Compromising less-secure systems by guessing passwords or using such password recovery systems to gain access to a computer without authorization can amount to a crime under the Computer Fraud and Abuse Act codified in 18 U.S.C. § 1030. In 2008, a hacker used the password recovery system to gain unauthorized access to the personal email account of then-Vice Presidential candidate Sarah Palin. He was charged and convicted of violating 18 U.S.C. § 1030(a)(2).⁴⁴

Passwords are also vulnerable when an attacker compromises a site and obtains the database of user passwords. In the worst case, those passwords are not encrypted and can be read directly. In other cases, the passwords are protected by hashing, hiding the original password. An attacker who acquires the hashes can try computing the hash of many different possible passwords to see if they match any of the stolen hashes; if they match, the attacker then knows the password. Using a number of graphics processing units, attackers can create billions or trillions of guesses per second. Passwords that are not sufficiently complex are thus easily obtained. There is also an attack against hashed passwords and other forms of encryption called rainbow tables. This is a method of pre computing many possible passwords or keys, which are stored in a table, reducing the time required to break passwords using brute force at the expense of the storage and pre computation.

Security experts recommend the use of a password manager. This is software that creates unique and complex passwords for every account. A user can create

44. Press Release, U.S. Dept. of Justice, Tennessee Man Sentenced for Illegally Accessing Former Governor Sarah Palin's E-Mail Account and Obstruction of Justice (Nov. 12, 2010), <https://perma.cc/XPH7-EBH3>.

and remember one complex password and use that to access all the stored passwords.

What You Have (Physical Tokens)

Possession of a unique item can also identify a user. Most often, these are used in conjunction with a password so that a stolen item alone is not sufficient for identification. Common mechanisms require the device owner to provide a response to some challenge, to provide a limited-time token to log in, or to perform a computation with a cryptographic key.

Mobile phones often are used for challenge response authentication. When logging into a site, the user might be sent a text message and be required to enter the code that it contains to prove possession of the device. There are similar programs, such as Google Authenticator, that run on a phone that provide a time-based code that changes frequently; before mobile phones were common, employers often provided small, dedicated devices that displayed this changing code on a small screen to provide the same functionality. Other devices include those that contain a cryptographic key. These include USB devices or smart cards that are inserted into a computer and that perform operations on the key embedded in the device.

What You Are (Biometrics)

Users can be identified by their physical characteristics. This is called biometric identification. There are a wide variety of measurements and observations about a person that can be made to determine if they are someone who has been seen by the system before and uniquely identified on that basis. The most common methods are facial recognition and fingerprints, but many others can be used as well, including retinal or iris scans, voice recognition, and hand shape. Some systems include a liveness check to ensure that the user is alive; this limits the ability of an attacker to murder the user or steal a body part for authentication, or to use a photo or fake finger to enter the system.

Biometrics is a detection problem. The system is trying to determine if a particular user is someone known to the system. As such, it does have a false positive and false negative rate, though modern systems can be very accurate. Biometrics also have additional error rates, known as failure to capture and failure to enroll. In the first, the sensor sometimes does not obtain adequate information to make a determination. In the second, it is possible that some users may be unable to use the system. As an example, it is not rare that some people are unable to use fingerprint recognition. People who have no easily measurable fingerprint ridges either owing to physiology, accident, or manual labor can fail to register on the system.

Requiring subjects to unlock devices using biometric security is considered distinct from requiring disclosure of passwords. Courts have generally allowed law enforcement to require individuals to provide biometric identifiers such as fingerprints or facial recognition to unlock a device, as they are considered physical evidence rather than testimonial evidence.⁴⁵

How Intrusions Happen

It is unfortunately common for computer security systems to be compromised and data to be accessed, modified, or stolen. A variety of actors take part in this process, each with their own motivations, goals, and abilities that affect how a breach occurs.

Types of Attackers

In the simplest case, security breaches happen because of mistakes by well-meaning but perhaps ignorant users or workers. They do not intend to create a problem, but sometimes can be enticed to do so by an attacker who uses social engineering. This is a technique that exploits human interaction to get a user to provide access to systems or information through fraud. Technical attackers have more sophisticated tools. One aspect of computer intrusion is that an expert can determine how to exploit a flaw in software and write a program to do so; one doesn't need to be an expert to run that program. The experts who are able to write tools to exploit flaws are often referred to as system crackers or hackers. This activity is not inherently illegal unless those tools are used to commit crimes. A black hat hacker typically has malicious intent and wants to break into systems to steal information or resources, or to extort the system owner, and may be involved with organized crime. Conversely, a white hat hacker finds exploits but notifies the software developer and does not exploit them; many companies have a bug bounty that pays rewards for these exploits so they can be fixed before they are a problem. In between these is a gray hat hacker who might not always act ethically; they might sell an exploit to others to use though they do not do so themselves. There is a large market for exploits. At the time of writing, a remotely exploitable but undetectable attack on an iPhone that gave the highest level of access was worth over \$1.5 million; it is unlikely that a bug bounty from Apple would prove as lucrative.

While companies like Google have dedicated white-hat teams, many of the best resourced attackers are those that are part of nation-states. Most large nations

45. *State v. Diamond*, 905 N.W.2d 870 (Minn. 2018).

have their own well-funded intelligence services that seek to penetrate other nations' systems for intelligence gathering.⁴⁶ These services have capabilities beyond any commercial organization, and they can exploit a variety of attacks against hardware and software in coordinated and sophisticated ways. Often, they gain access to computers and networks and can keep a presence in there even if they are discovered; these are referred to as advanced persistent threats. Smaller nations turn to commercial suppliers of online intelligence tools; these tools are still often very effective, particularly against citizens being surveilled. More recently, tech companies have started suing these companies.⁴⁷

The most damaging attackers might be insiders. These are users who are given some trust within an organization and who have access to some or all internal systems, then abuse that trust to conduct attacks. Sometimes they do so after being recruited by other nations; other times they do so for financial gain or because they are disgruntled.

Types of Intrusions

There are a wide variety of ways to intrude on a computer system, most of which are outside the scope of this guide. As a general overview, though, most attacks occur through stolen authentication credentials or technical flaws in systems or software.

Attackers gain access to user credentials in a variety of ways. The easiest is to trick the user into providing credentials by getting them to visit a fake website that will capture what the user enters; hence the large volume of phishing emails and texts that are sent every day. At other times, attackers are able to find leaked user passwords that are posted online and then try them at other websites. Sometimes other credentials, such as for cloud computing services, end up stored online but without proper confidentiality settings and are later discovered and abused.

Attackers are also able to gain access to systems by exploiting flaws in the running code. A general approach for this is for the attacker to find a flaw in the system in which attacker input gets interpreted as computer code and executed, or to find input that causes an unanticipated effect. The techniques for this vary depending on what kind of software is running and what input the attacker can provide.

Locating and identifying attackers who conduct remote attacks over the internet can be challenging or impossible. These attackers can try to hide their

46. U.S. Dept. of Justice, "Two Chinese Hackers Working with the Ministry of State Security Charged with Global Computer Intrusion Campaign Targeting Intellectual Property and Confidential Business Information, Including COVID-19 Research" (Off. of Pub. Affs., Nov. 27, 2023), <https://perma.cc/MJS6-QMD7>.

47. Apple Inc. v. NSO Grp. Techs. Ltd., 5:21-cv-9078.

location. They might use an anonymous network service like Tor, or they might intrude on other computers and use those to direct attack traffic through; the eventual victim only sees traffic coming from these intermediate computers and not from the computer originating the attack. Botnets, described in the malware section below, are used to facilitate this.

Legally Authorized Intrusions

Not all computer intrusions are illegal. The law criminalizing unauthorized access to a protected computer includes an exception for authorized law enforcement activity (*see* 18 U.S.C. § 1030(f)). However, this statutory exception does not address Fourth Amendment protections. Thus, when law enforcement seeks to gain access to a computer in this manner to collect information protected by the Fourth Amendment, they often obtain a search warrant. When executing this type of search warrant, law enforcement might overcome the computer's security using techniques similar to those a hacker might use, just as when law enforcement executes a court-authorized search of a home utilizing lock-picking tools similar to those used by a thief. Law enforcement might identify a technical flaw to which the systems are vulnerable; create or purchase an exploit that can take advantage of the vulnerability; deploy the exploit to gain access to the systems; then deploy a payload that will gather the authorized evidence; and collect that evidence at a law enforcement-controlled computer elsewhere on the internet.

When the location of the target computer is concealed through technological means, such as through obfuscation provided by an anonymous network, Federal Rule of Criminal Procedure 41 provides a court in the district where activities related to the crime may have occurred the jurisdiction to issue a warrant authorizing law enforcement to remotely search a computer located outside the court's district. Criminal Rule 41(6)(A) was added to the federal rule governing search warrants in 2016 and eliminates the jurisdictional challenges that occurred across the nation when the FBI obtained a single search warrant in the Eastern District of Virginia for the deployment of code from a seized child exploitation website called "Playpen" operating on the Tor network. Under this search warrant, the code was deployed from the government-controlled "Playpen" website to the computers visiting the site. This code conducted a limited search and forced the user's computer to reveal its true IP address to the FBI. Prior to this remote search provision in Rule 41, it was unsettled which court had the appropriate jurisdiction to issue such a search warrant. This question is now clearly resolved by this remote access provision to Rule 41.⁴⁸

48. Some of the federal cases addressing this jurisdictional question include *United States v. Henderson*, 906 F.3d 1109 (9th Cir. 2018); *United States v. Horton*, 863 F.3d 1041 (8th Cir. 2017);

Malware

Malware is malicious software that is designed to operate without the knowledge or consent of the computer user. It is often useful to differentiate how malware gets spread or placed on a computer from what it does once it is there. There are a wide variety of mechanisms used to get malware installed, and malware can have a variety of effects, including targeted ones that are specific to a victim. For descriptions of various types and components of malware, see the Glossary of Terms.

It is worthwhile to note that there is a market for access to computers infected with malware. Attackers will create a botnet by infecting many computers with malware that allows continuing access—these infected computers are called bots—and then install other malware on behalf of others who pay for access to the compromised computers. For example, in one case a Russian-born cybercriminal claimed on online forums that he could control up to 500,000 infected computers at one time.⁴⁹ When the threat is large and difficult for individuals to mitigate, courts have approved government efforts to enter compromised machines and remove the malware that was part of the botnet.⁵⁰

It is common for computer users who are accused of a computer-related crime to claim that they are innocent of it, and that any evidence is a result of malware installed by some outside actor. This is not an impossible occurrence, especially for a nontechnical user targeted by someone more technically sophisticated. In these cases, however, a forensic examiner should be able to determine if this is likely to have occurred by looking for evidence of malware or correlating user activity with evidence about the time of the alleged criminal activity. Often, this defense is undermined by robust evidence of user interaction with the computer or evidence during the time in question.

A failure to implement reasonable security measures to protect against a foreseeable risk can lead to civil liability under a claim of negligence.⁵¹ For a plaintiff's suit to survive a motion to dismiss, the court must consider if the complaint failed to establish one or more of the elements of a negligence claim, including causation between the defendants.⁵² If the complaint is found insufficient, courts will often grant leave to amend the complaint.

United States v. Werdene, 883 F.3d 204 (3d Cir. 2018); *United States v. Moorehead*, 912 F.3d 963 (6th Cir. 2019).

49. Press Release, U.S. Dept. of Justice, Russian-Born Cybercriminal Sentenced to Over Nine Years in Prison (July 12, 2017), <https://perma.cc/6EZG-L69P>.

50. Press Release, U.S. Dept. of Justice, Justice Department Announces Court-Authorized Disruption of Botnet Controlled by the Russian Federation's Main Intelligence Directorate (GRU) (Apr. 6, 2022), <https://perma.cc/PNY7-FQV5>.

51. See *In re Am. Med. Collection Agency Inc. Customer Data Sec. Breach Litig.*, 2021 WL 5937742 (D.N.J.); *In re Equifax Inc., Customer Data Sec. Breach Litig.*, 362 F. Supp. 3d 1295 (N.D. Ga. 2019); *In re Home Depot, Inc., Customer Data Sec. Breach Litig.*, 2016 WL 2897520 (N.D. Ga.).

52. See *Aspen Am. Ins. Co., et al. v. Blackbaud, Inc.*, 624 F. Supp. 3d 982 (N.D. Ind. 2022); *In re Sony Gaming Networks & Customer Data Sec. Breach Litig.*, 996 F. Supp. 2d 942 (S.D. Cal.

Forensics and Digital Evidence

As stated in the Introduction, many legal issues involve facts and evidence related to software, computers, networks, and digital information, all of which have become ubiquitous in our personal lives, as well as in commercial and government processes, products, and services. For example, digital evidence is an aspect of almost every criminal investigation.⁵³

Forensics is the scientific investigation and formal presentation of evidence that supports or refutes an investigative hypothesis that explains an event of interest. When the investigator follows a repeatable, structured process for gathering evidence and uses strong inductive reasoning, forensics is a science rather than merely a set of techniques that recover data. A successful hypothesis has supporting evidence that explains allegations as a series of actions, events, identities, or intentions, and does not have strong countervailing evidence. Digital forensics is focused on evidence related to computer and network systems.

Evidence should be gathered following standard techniques that are defined ahead of the crime and accepted by the scientific community. The investigator attempts to identify sources of possible digital data, copy such data, and preserve its integrity; extract relevant evidence from the collected data; and report the results of the investigation.⁵⁴ These stages transform identified artifacts into evidence that speaks to facts, or that link data to a scene or person through individualization or can speak to the intentionality of a person.

Recovery of Digital Evidence

A computer comprises many systems: file systems, databases, networking, and more. Each system is designed to present a simple view of a complex set of interactions.

For example, the networking system hides a large amount of work that goes into presenting a web page. Similarly, the file system presents your files to you without the details of how files are actually managed. The operations that are hidden from the user by each system are designed to operate efficiently, and often efficiency means leaving artifacts behind.

This is particularly true when files are deleted, because the quickest mechanism for the file system is to simply mark the file as deleted, then to later

2014); *Heritage Valley Health Sys., Inc. v. Nuance Communications*, 479 F. Supp. 3d 175 (W.D. Penn. 2020) (malware attack sponsored by Russian military—NotPetya).

53. David B. Muhlhausen, Report to Congress: Needs assessment of forensic laboratories and medical examiner/coroner offices (National Institute of Justice, May 2020).

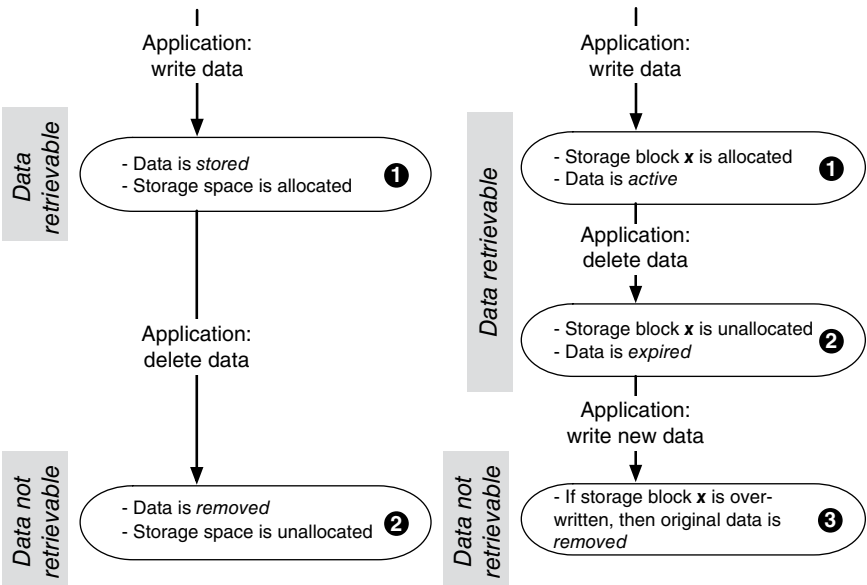
54. Keith Inman & Norah Rudin, *The origin of evidence*, 126 *Forensic Sci. Int'l* 11–16 (Mar. 2002), [https://doi.org/10.1016/s0379-0738\(02\)00031-2](https://doi.org/10.1016/s0379-0738(02)00031-2).

overwrite the file when space is needed. This process is similar to writing with a pen. It is easier to cross out mistakes from sentences in a letter than it is to re-copy the entire letter’s text onto a new sheet of paper, but the old writing might still be legible.

The file system presents to the user a simplified logical view of the file system that doesn’t include all the information available to the file system itself. In fact, the goal of every system is to present a complex service as something simpler via an interface. One analogy to this process is dining at a restaurant. The menu and the waiter presented to the customer are an interface that can be used to order food from the kitchen. Most kitchens are chaotic and messy, yet the plates that arrive at the table don’t show it. It is the goal of the waiter to present such an illusion, and any computer interface is no different.

Figure 9 illustrates the user’s view of the file system. The interface allows the user to create files, and storage space is allocated accordingly. The user can issue a command to delete the file, and storage space is then unallocated. File modifications can occur through overwriting the same storage space with new data, finding space for additional data, or deleting the existing file and then creating a new file with the same name. Users have no information or influence regarding where a file is stored on the storage medium—that is a complication handled by the file system.

Figure 9. Left: A user’s view of the file system interface. Right: The true life cycle of data and files in a file system.



For forensic investigators, the life of data in storage is not so simple. Figure 9 illustrates the internal life cycle of files, data, and storage on a file system, beginning with data given to the file system by some application, such as a word processor. In Step 1, a file is written to a specific allocated block *x*, and we say that it contains active data. When the user or application issues the command to delete the data, the process moves to Step 2, where the storage block is unallocated. The data remains present in storage, but we say it is expired; that is, the data cannot be recovered through the file system's interface to the user, but it is retrievable by an investigator with direct access to the disk. Finally, once the same block is allocated to new data, the overwritten portions of the old data are removed and unavailable to the investigator. For decades, forensic experts have identified and examined many other types of digital evidence available from computer systems, including desktops and laptops, smartphones, wearables, IOT devices, cloud systems, and more. The same principles as above apply: data are stored on these devices because the data's presence increases usability, performance, and security; data are often copied between and stored on multiple systems because of interactions among parties; and such data often remain available for recovery even after deletion.

Relatedly, it is notable that courts have serious concerns when digital evidence is intentionally destroyed. For example, in *Burris v. JP Morgan Chase & Co.*,⁵⁵ a whistleblower protection case was dismissed with prejudice because of a plaintiff's systematic destruction of electronically stored information (ESI). The court appointed a computer forensic examiner who reviewed the plaintiff's electronic devices and discovered evidence that ESI had been systematically deleted. In addition, two software tools that advertised their ability to overwrite free space, making it impossible to recover deleted data, were found on plaintiff's electronic devices.

Computer Science from the Perspective of Federal Rule of Evidence 702

When considering experts for testifying on computer science matters, the factors elucidated by the Supreme Court in *Daubert v. Merrell Dow Pharmaceuticals, Inc.*⁵⁶ and its related cases⁵⁷ apply well. As the Supreme Court explained, a trial court's evaluation of the underlying reasoning supporting an expert's testimony and the validity of the scientific methodology they use should be "flexible . . . and its focus

55. 566 F. Supp. 3d 995 (D. Ariz. 2021).

56. *Daubert v. Merrell Dow Pharms., Inc.*, 509 U.S. 579 (1993).

57. See also *General Electric Co. v. Joiner*, 522 U.S. 136 (1997); *Kumho Tire Co., Ltd. v. Carmichael*, 526 U.S. 137 (1999).

must be solely on principles and methodology, not on the conclusions that they generate.” The factors established in *Daubert* to be used in a court’s inquiry should be inclusive of:

- “*whether the theory or technique in question can be (and has been) tested.*” If a method is testable, then there are demonstrable, controlled scenarios where the test can result in positive and negative results.
- “*whether it has been subjected to peer review and publication.*” Peer review is the process by which a paper is evaluated by qualified scientists working in the same field as the subject of a submitted paper. In many fields outside of computer science, journals are peer reviewed and conferences are not. However, in computer science, not only are conferences almost always peer reviewed, but they are also often considered the most prestigious venues for publishing (implying a paper’s results have been more carefully evaluated by the reviewers).
- “*its known or potential error rate and the existence and maintenance of standards controlling its operation.*” Ideally the methodology supporting testimony has a quantified error rate based on data independent of the case in front of the court. Notably, there is no requirement that the error rate be at a specific level, only that it be known to the court.
- “*whether it has attracted widespread acceptance within a relevant scientific community.*” Methodologies are often based on known approaches that already appear in textbooks, previous peer-reviewed publications, or reports on best practices from established organizations. Evidence and facts should be based on the results of practices and protocols that are written down in lab manuals ahead of an investigation. This approach often has the advantage of avoiding the appearance of bias. The training and methodologies used by practitioners and examiners should have been developed by experts who have applied a scientific methodology.

In computer science, two international societies stand out as having a strong track record of overseeing high-quality conferences and journals: the Association for Computing Machinery (ACM) and the Institute of Electrical and Electronics Engineers (IEEE). Other high-quality publication venues and societies certainly exist.

These factors were codified in a series of amendments to Federal Rule of Evidence 702 in an effort to clarify their application.⁵⁸ Assessment of admissibility of computer science evidence follows an analysis that includes the *Daubert* questions discussed above: Are the methods testable and have they been tested? Has peer review been employed? What is the quantified error rate? Has there been

58. For a discussion of Federal Rule of Evidence 702, see Liesa L. Richter and Daniel J. Capra, *The Admissibility of Expert Testimony*, in this manual.

widespread acceptance? Whether a particular witness is qualified per the expert's scientific, technical, or other specialized knowledge—codified as Federal Rule of Evidence 702(a)—is discussed presently.

Experts in Computer Science

Overall, determining if an expert is qualified should be based on criteria specific to the case at hand. For cases involving computers and computer science, an appropriate expert is one who has experience and training relevant to the particular aspect of computer science involved. By analogy, consider a case that involves accusations of malpractice related to heart surgery—surely a podiatrist is the wrong choice to serve as an expert witness. Like medicine, computer science is a broad field and experts have specialties. These specialties are myriad: networking, security, machine learning, databases, algorithms, to name a few. Moreover, the field is both a science and an applied discipline. A person with years of research experience may not necessarily be able to speak to the complexities of a deployed production system; and likewise, years of applied experience does not necessarily imply an understanding of scientific fundamentals, nor a deep and rich understanding of the limits and consequences of applying techniques and methods.

However, speaking broadly, there are many ways to evaluate whether a particular person is a qualified expert in computers or computer science. These include asking if the candidate has academic degrees in computer science or computer engineering from well-known universities and colleges, awards for high-quality papers, advanced member grade status from professional societies (such as being a “Fellow” of the ACM or IEEE), a high count of publications in peer-reviewed conferences and journals, and a high count of citations to those papers (though citation counts vary in magnitude across subfields of computer science). Just as notable is a record of years of experience in a computer science–related role with reputable companies, government agencies, and other institutions.

Trial courts are considered the gatekeepers of expert testimony, per *General Electric v. Joiner*. For example, in *Resnick v. United States*,⁵⁹ the court found there was no right to call a computer expert to rebut the government's expert witness. In that case, the defendant missed the deadline to file notice of an intended computer expert. The court found the defense counsel's questioning of the government computer expert sufficient and not evidence of ineffective assistance of counsel.

59. 7 F.4th 611 (7th Cir. 2021).

The Quality of Forensics and Digital Evidence

Forensic investigations are at their strongest when conclusions are drawn via inductive reasoning. In this type of reasoning, a generalization or model of what is expected to occur is created based on what has been observed in repeated observations. Ideally, these observations have been made independent of the facts of the case. Other aspects related to *Daubert* should also be present, as listed above: the study of observations should make use of well-known methods, and the conclusions should be published in a peer-reviewed scientific venue. The reasoning should involve a falsifiable hypothesis, and the observations should provide for a known, quantified error rate.

Inductive reasoning is strongest when hypotheses are verified by independent parties. Investigators rely heavily on validation studies that perform repeated and precise tests on equipment and software to determine what can be said with assurance about evidence. Validation reports are published by government agencies, such as the National Institute of Standards and Technology (NIST),⁶⁰ by industry and academic researchers in peer-reviewed journals and conference proceedings, and by professionals who engage in testing of their tools.

Direct evidence proves a fact without inference, such as a voluntary confession would provide. Digital artifacts are often circumstantial evidence of an event or fact, but that status should not be viewed pejoratively. Direct evidence from witnesses is not always reliable. People do not have perfect perception or memory. People have personal biases and can be paid or otherwise motivated to give false testimony. Confessions can be coerced in not very subtle ways. In fact, most evidence at a crime scene is indirect evidence. For any crime for which there are not witnesses, the case must be circumstantial.

Digital evidence is typically modifiable.⁶¹ For that reason, one might conclude that digital evidence should not be considered as absolute fact when it is found—but in fact it is no weaker than other types of circumstantial evidence. DNA evidence has a similar problem: it's easy to plant DNA evidence at a scene—a few skin cells will do it.

Further, indirect evidence can be stronger than direct evidence if there is other corroborating evidence—a notion that Locard's Exchange Principle speaks to. For example, let's say that it has been alleged that John committed a crime against Jane, and John claims to not know her at all. Investigators find that John's web browser has a history of pages he has visited recently, including the text and images from those pages. Jane's public webpage is found in that history cache, and it is used as indirect evidence that he knew Jane before the crime took place. John's

60. NIST, Computer Security Resource Center, <https://perma.cc/4UXM-B6E8>.

61. Forensic practitioners record and report cryptographic hashes of collected material to help identify any modified data and as a best practice should take care not to accidentally modify any data.

browser will record when exactly John last visited Jane's page, and such facts can be corroborated by examining the web server that hosts Jane's webpage. Furthermore, other logs at John's internet service provider may be able to confirm indirectly that his computer was connected to the internet at the time the page was viewed. Logs from John's email server may indicate he checked or sent email at the time when the webpage was retrieved; if he admits to keeping his account and password secret from others, then the email server logs indicate he was at the keyboard at the time.

E-Discovery

Digital forensics is often part of an adversarial investigative process in which an investigative hypothesis about a user's actions is confirmed or refuted based on evidence collected from electronic devices. The assumption is that an individual under investigation might try and conceal data about their actions and authorities need to act to collect data before it is lost or destroyed.

E-discovery differs from forensics in that it is part of a civil discovery process. While forensics techniques might be used in some specific circumstances, such as making a preservation request for a particular user's devices to obtain data and metadata from them, in general e-discovery will require parties in a civil action to produce relevant material they already have in an electronic form to allow faster and less expensive review. Normal discovery rules apply; each party is bound to produce material responsive to the case, but each party decides what materials are responsive rather than the other party having access to all data and making that determination, which can clearly be intrusive.

E-discovery allows for faster and cheaper review than paper. Documents can be searched using a text retrieval system that allows searching for terms. Additionally, supplying documents electronically allows the document metadata to be used. This can include dates when documents were created; dates and times for emails and other messages; and other relevant information that can be searched as well.

There are a variety of commercial solutions for uploading, managing, and reviewing documents. The Sedona Conference has published documents to help guide judges through the e-discovery process.⁶²

62. The Sedona Conference, Resources for the Judiciary, <https://perma.cc/Z3AJ-UZMC>.

Glossary of Terms

This section contains definitions of terms from the text and others that might be legally relevant.

access point (AP). A hardware device that provides wireless internet access to nearby devices over Wi-Fi.

advanced encryption standard (AES). The U.S. government standard for shared-key or symmetrical encryption for all except classified information. This protocol replaced DES in 2005. It allows use of keys of 128, 192, and 256 bits, which are large enough to resist brute-force attacks.

advanced persistent threats (APT). A sophisticated and targeted cyberattack carried out by highly skilled and well-resourced attackers who seek to gain unauthorized access to a network or system for an extended period of time.

adware. A term for legitimate software that is supported by advertisers; can also be malware that illicitly serves advertising to computer users. Unscrupulous attackers get paid by an often-unsuspecting advertiser for what appear to be legitimate ad views. Instead, the ads either replace legitimate ads on a webpage, are added to webpages, or are never shown to a user but are charged for.

algorithm. An ordered series of steps used to accomplish a task. A program is a representation of an algorithm that a computer can execute.

application. See program.

authentication. The process of verifying a user's identity in a secure way to provide access to a system.

availability. The security design goal of ensuring that data and systems are available when needed.

backdoor. A mechanism that is installed to allow future access to the computer that bypasses the normal login mechanism. This is used either to access the computer later in case the installed malware is removed, to gain entry to the machine while bypassing the normal logging mechanism, or both.

base rate fallacy. A problem with detection systems in cases where the thing being sought is rare compared to the number of items to examine; it results in most alerts being false positives, degrading the effectiveness of the system.

Big-O or Big-Ω. Terms used by computer scientists to measure the efficiency of an algorithm in terms of the number of computational steps needed to solve a problem given an input of size n , the number of input items. Big-O ("big oh") notation describes the worst-case running time for an algorithm; this is the most steps it will ever take to compute the algorithm. Similarly, Big-Ω ("big omega") notation is the best possible case for the number of steps needed and supplies a lower bound.

binary encoding. The representation of a value (a number or character) in base 2, i.e., using only ones and zeros.

biometric. A measurement of some physical aspect of a person used in authentication processes to determine identity.

birthday attack. This term is a reference to a type of cryptographic attack where a match is found between any two pairs of items in a set. This type of match is easier to find than fixing one item in the set and looking for a match among the remaining items. The reason the birthday attack is easier is that there are more pairs of items in the set than there are remaining items after selecting one; for example, in a set of 10 items, there are 45 pairs; but if we select one item, only 9 remain. The name of the attack refers to the high chance of finding two persons with the same birthday from among a relatively small group.

bit. A single binary digit that is a 1 or 0.

black hat. A computer attacker/hacker who has malicious or criminal intent.

blockchain. A distributed and decentralized digital ledger that records transactions in a secure and tamper-evident way.

boot sector. The first sector of a storage device, such as an SSD, that contains information about the file system including where code to boot the operating system lies.

bots/botnet. Attackers sometimes take over computers to be part of a wider distributed network that participates in a variety of activities; the computers that are forced to participate are known as bots and the entire group a botnet. Botnets are often used for sending spam emails or text messages; they can also be used as part of a distributed denial-of-service attack, where the bots send large amounts of garbage traffic that overwhelms the internet or a targeted computer.

browser hijacker. Software that infects a user's web browser to automatically click on ads or to redirect traffic to particular webpages that contain ads.

bugs. Errors or defects in software code that can cause unexpected or incorrect behavior in a program.

buses. Electronic interconnections in a computer that move data between its components

byte. Eight bits together; typically, the smallest commonly used measure of data. Measuring bytes commonly includes prefixes, like kilo- for 1,000, mega- for 1 million, giga- for 1 billion, and tera- for 1 trillion.

caches. Small but fast repositories of memory used to store data and speed computation in a CPU. (A CPU cache is just one type, and other caches appear elsewhere in a computer system.)

cell site location information (CSLI). A system that can be used to triangulate the location of a cell phone based on receiving a signal at several different cell towers.

central processing unit (CPU). The primary component of a computer that performs the majority of the processing and controls the other components.

certificate. A cryptographic credential used to verify the identity of websites or the creators of software. A root certificate is a credential that comes with or is installed in an operating system or internet browser that is used to verify the ownership of other certificates.

ciphertext. Text that has been encrypted to be unreadable without the appropriate key.

client-server. A computing communications model in which a centralized system, known as a server, interacts with many subordinate clients.

closed-source. A type of source code that is maintained privately by the creator; contrast with open-source.

cloud computing. A large-scale computing model in which programs are run on computers elsewhere in the network; often these computers are rented out by large cloud providers.

code. See source code.

compiler. A computer program that takes human-readable computer source code and translates it into a binary form executable by a computer; this binary form is referred to as having been compiled.

computational core. The portion of a CPU that executes a program; many CPUs now contain multiple cores.

confidentiality. The security goal of keeping information secret from those who are not authorized to access it.

cookie. A small piece of information recorded in a web browser that is set by and shared with individual websites. Used for authorization, advertising, and to track the user's actions across a site.

copyleft. A type of licensing agreement used for software that allows for the free use, modification, and distribution of the work under certain conditions. The copyleft license requires that any modified or derived work from the original must also be licensed under the same terms as the original work.

cores. See computational core.

crackers (or hackers). Technologically sophisticated attackers who are able to breach computer security measures.

crypto jacking. Many cryptocurrencies require that computers called miners participate to allow transactions to proceed; miners occasionally get rewarded for their work. Attackers install malware that acts as a miner on the user's

computer. This is called crypto jacking, and it allows the attacker to earn some cryptocurrency at the user's expense.

cryptographic hash. An algorithm that takes input of any size and produces a fixed-size output that represents a unique fingerprint of the original input. This output is also referred to as a hash value, message digest, or checksum.

cryptographic signature. Data that authenticate both the author and contents of a message in a manner that is mathematically provable to someone who knows the following: the algorithm used to sign the message; the message; and a cryptographic key created by the signing author.

darknets. A type of network that is intentionally hidden and can only be accessed using special software; often access is limited to authorized users.

data center. A physical location that hosts a large number of computers. Typically, these locations provide physical security, electricity, cooling, and network access.

database. A specialized system that stores data in a relational format and offers a method of querying and processing the stored data.

decompilation. A method of transforming a compiled program back into human-readable source code. Often the decompilation process is not perfect, as information is lost during compilation.

decryption. The process of converting ciphertext into plaintext by applying an algorithm and the appropriate key.

denial of service (DOS). An attack against computer or network systems with the intent of making them unusable or unreachable.

dictionary attack. An attack against password systems where the attacker encrypts large numbers of possible passwords to see if they match an unknown encrypted or hashed password.

digital signature. A cryptographic technique that provides authenticity, integrity, and non repudiation to a digital document by encrypting the document or its hash with the signer's private key.

drive-by downloads and **watering hole attacks.** A drive-by download exploits internet web browsers that have errors that allow websites to force installation of software on computers that visit the site. To find victims, attackers can choose sites that are close in name to a real site and hope that users make a typo when entering the address. They might also place ads that carry malware on internet advertising networks. In a watering hole attack, an attacker might create or compromise a website that appeals to some population they want to attack, like systems administrators, so that malware will be more likely to reach that group.

dropper. The most common item installed through malware, a dropper is software that allows the attacker to install other software. There are underground

markets where attackers pay to access computers that are already infected with droppers so they can install additional malware.

electronic communication services (ECS). Any service that provides users the ability to send or receive wire (voice) or electronic communications. *See* 18 U.S.C. § 2510(15). Some examples include email providers, chat sites, websites that provide direct messaging services, and texting apps.

encryption. An algorithm that uses a key to transform data so that it appears random and protects confidentiality.

end user license agreements (EULAs). Legal agreements between software providers and users that outline the terms and conditions of use.

Ethernet. A hardware computer network protocol used to transfer data over wired local area networks (LANs).

executable. An executable file containing binary instructions that can be run directly by a computer's processor.

file system. The portion of the operating system that structures data and organizes files and directories on a computer's storage device.

flash memory. Non volatile computer memory that can be electronically erased and reprogrammed.

free and open source (FOSS). Software that is licensed in a way that allows users to freely use, modify, and distribute the source code.

Freenet. A peer-to-peer network designed to provide secure and anonymous file sharing.

functions (methods, or procedures). Self-contained blocks of code that perform some algorithm and can be called from other parts of a program.

geolocation. The process of determining the physical location of a device or user, often based on IP address; this can be inaccurate.

GNU public license (GPL). A FOSS license that grants users the freedom to use, modify, and distribute software as long as any derivative works are also licensed under the GPL.

graphics processor unit (GPU). A computer component with many weaker processing cores, primarily useful for rendering images and video but also useful for machine learning and cryptocurrency.

hackers. *See* crackers.

hard disk. A storage device used to store and retrieve digital information using magnetic technology.

hash. *See* cryptographic hash.

International Mobile Equipment Identity (IMEI). A unique identifier assigned to mobile phones and other cellular devices.

internet of things (IOT). Physical devices with sensors, software, and network connectivity that exchange data with other devices and systems over the internet.

internet protocol (IP). The portion of the internet software and protocol stack that handles delivery between computers on different networks.

internet service providers (ISP). Companies that provide customers with access to the internet.

interpreted languages. Programming languages that are executed by an interpreter as they run rather than being compiled into machine code, which is run separately.

interpreter. A program that reads and executes code written in an interpreted programming language.

IP address. A numerical identifier assigned to each device connected to the internet; this address may not be unique to a particular system owing to NAT and other mechanisms.

keylogger. A form of spyware that monitors what the user types, and sends it to the attacker or saves it in a file for the attacker to recover later. This can be used to capture passwords, emails, or other things that the user types.

library. A collection of prewritten functions that can be used by a program to perform common tasks.

link layer. The portion of communication software responsible for communication between devices on a local area network (LAN), such as devices on Ethernet or Wi-Fi.

machine learning. An area of artificial intelligence that enables computers to learn from and make decisions based on data without needing a human creating a particular algorithm.

malware. Software intended to cause harm to or violate the security of a computer system.

medium access control (MAC). A portion of the link layer in the networking stack that controls access to the physical medium of a network.

memory (non volatile). Computer memory that retains data even when the power is turned off, such as flash memory or hard disk drives.

memory (random access, volatile). Computer memory that can be accessed randomly and quickly, but loses its data when the power is turned off (such as RAM).

metadata. Data that describes other data, such as the author, date, and file type of a document.

mining. The process of verifying and recording transactions on a cryptocurrency blockchain by solving complex mathematical problems with dedicated computers.

motherboard. The main printed circuitboard in a computer that connects all of the other components and peripherals.

network address translation (NAT). A technique used to remap one IP address space into another by modifying network address information in the IP header of packets while they are in transit across a traffic routing device.

network interface. The physical or virtual connection point between a computer and a network.

network layer. The layer of the networking stack responsible for routing packets between networks.

open source. See free and open source (FOSS).

operating systems. Software that manages computer hardware resources and provides common services for computer programs as well as a user interface.

optical media. Storage media that uses optical technology to read and write data, such as CDs, DVDs, and Blu-ray discs.

packet capture (pcap). A file format used by Wireshark and other programs to capture and store network traffic for analysis and debugging purposes.

phishing and spear phishing. The act of an attacker who wishes to target a specific individual or group by sending an email, text, or other messages that contain a link to malware or to a site that can install it; when sent widely, this is referred to as phishing. Spear phishing is a more targeted technique in which the attacker learns about the victim through social media or web presence and then forges email from someone the victim would trust. Spear phishing is also frequently used in computer fraud—for example, altering an account number where funds are to be transferred during a transaction so that the funds go to the attacker instead of the intended recipient.

physical layer. The lowest layer of a network software stack that handles the physical transmission of data over a communication channel.

port numbers. Numbers in internet connections used to associate a network connection with a particular process that is sending or receiving data.

program. Computer executable instructions that perform a specific task on a computer system.

programming language. A formal language used to write human-readable algorithms that can, through interpretation or compilation, later be executed by a computer to perform a specific task.

public-private key pair. A set of cryptographic keys that are used to encrypt and decrypt messages and to verify the digital signatures of digital certificates.

rainbow tables. A method of pre computing and storing many possible passwords or keys to reduce the time required to break a password or key compared to brute force computation alone.

random access memory (RAM). Circuitry that holds working data while the computer is running. RAM is volatile, meaning it needs power to continue storing information.

ransomware. Malicious software that encrypts user files, then supplies an apparent route to recover them through paying the attacker—this type of software has become a significant risk. Organized criminal gangs have started targeting larger organizations that are dependent on data, such as banks, universities, and hospitals; cryptocurrency has provided a mechanism for criminals to receive their ransom while concealing their identity and location. One method deployed by law enforcement to fight this criminal activity is to gain control over the servers and websites used to communicate and control its victims.

remote access tool. See spyware.

root certificates. A file containing the public key of a trusted authority. The public key is used to digitally sign the public key certificates of its customers, ensuring they are authentic.

root kit. Malware that subverts the normal operating system observability tools in order to hide the existence of the malware.

router. A specialized computer that connects and transfers data among multiple computers. The router is able to make decisions about which of several possible next “hops” the traffic should be sent to, ensuring that the shortest path to the destination is taken.

scareware. Attackers often try to frighten users into sending money with scareware. One form of scareware is software that produces fake antivirus warning messages that request the user pay for a software upgrade to remove a non-existent threat.

serverless computing. See cloud computing.

shared key. See symmetric key.

social engineering. A process by which attackers gain access and information by convincing a user to provide access to systems or information through fraud.

software. A set of instructions that can be executed by a computer.

solid state drive (SSD). A high-capacity, non-volatile storage device to retain data.

source code. The set of files written in a programming language that are used to create a program. Might also refer to a production of these files along with additional files relating to the design and construction of the program.

source code control. A system used by software engineers to track changes made over time to source code. Typically, the source code (and the log of the changes) is said to reside in a repository.

spear phishing. See phishing.

spyware. A general term for software that can monitor user actions and data on a computer. In terms of malware, this might be a remote access tool (or RAT) that provides complete remote control of the computer, including the ability to access files on the system, add files to the system, or activate a camera and/or microphone.

subnet. A smaller network that is part of a larger network.

subscriber identity module (SIM). A small, removable card used in mobile devices to identify and authenticate a subscriber on a cellular network. Newer phones use eSIMS, which are not physical but are stored as information on the phone.

symmetric key. An unguessable value shared as a secret by two participants exchanging encrypted data via a symmetric-key encryption algorithm. The same key is used to encrypt and decrypt.

tape. A data storage medium that uses magnetic tape to store digital information, most often for long-term data storage and backup purposes.

Tor Browser. Free, open-source software from the Tor Project that enables a user to join a network of other users volunteering as network relays. All traffic from the user is sent through the relays such that the website (or other destination) the user is contacting cannot link the received traffic to the user's real IP addresses. Tor Browser provides a service that is similar to a VPN but is more secure in some ways.

transport control protocol. The portion of the internet software and protocol stack that ensures data is delivered in a reliable manner from one computer on the internet to another.

Trojan or Trojan horse. A program that promises, appears to, or really does supply some useful functionality while secretly performing harmful actions.

USB (universal serial bus) drive. A portable storage device that uses flash memory to store data.

user datagram protocol (UDP). A transport layer protocol that provides an unreliable but low-latency network connection.

version control software. A program that serves as a repository for source code and other documents that is able to track and revert changes.

virtual machines (VMs). A software program that pretends to be computer hardware while running on a host's real hardware so that other programs or operating systems can be run while isolated from the host computer.

virtual private network (VPN). An encrypted connection that forwards data across other networks that can only be decrypted and read at the endpoint; this endpoint might forward the received data, making it appear as if the endpoint is where the initial computer resides.

- virus.** A program that can copy itself by inserting its code into other programs. The infected program then finds other programs that it can infect and adds the virus there. It is common for viruses to spread when files and programs are shared. Designers often take steps to obfuscate what the virus is doing. They might repack their code, which keeps the same functionality but changes the appearance of the code. They might encrypt much of the functionality and decrypt it only as needed. Both techniques and others can be used so that the virus is different on each infected system; this is an example of a polymorphic virus. Viruses also attempt to elude detection, possibly by starting before the operating system; boot sector viruses are an example of this approach.
- VPN concentrator.** The device a computer will connect to send data through a VPN.
- watering hole attack.** See drive-by download.
- white hat.** An ethical hacker who works to improve computer security (see also black hat).
- Wi-Fi.** A commercial term for a series of industry standards for short-range wireless communication among computers and mobile devices.
- Wireshark.** A packet capture tool used to monitor, log, and/or debug network traffic.
- worm.** A computer worm is a self-executing program that runs on its own, unlike a virus. A worm spreads across a network by taking advantage of security vulnerabilities on other network-connected systems. A worm can spread shockingly fast in some cases; it might be possible to infect all vulnerable systems worldwide in under 15 seconds in some cases. A major worm spreading is an uncommon and often newsworthy event as it typically causes large network congestion and outages.

Reference Guide on Artificial Intelligence

JAMES E. BAKER AND LAURIE N. HOBART

The Honorable James E. Baker, J.D., is the Director of the Syracuse University Institute for Security Policy and Law, Professor of Law at Syracuse University College of Law, Professor of Public Administration (by courtesy) at the Maxwell School of Citizenship and Public Affairs at Syracuse University, and a judge on the Data Protection Review Court.

Laurie N. Hobart, J.D., is Associate Teaching Professor at Syracuse University College of Law.

CONTENTS

Introduction, 1484

Roadmap, 1486

An Overview of AI, 1486

What Is AI?, 1486

A Brief History from Turing to Today: Waves of AI Development, 1488

Complex Algorithms, 1489

Computational Speed, 1490

Sensors, 1490

Data, 1491

Cloud Computing, 1491

Machine Learning, 1491

A Survey of AI Applications “Today,” 1494

What Is Machine Learning?, 1500

How Machines Are Trained to Learn: The Machine

Learning Life Cycle, 1501

Problem Specification, 1502

Data Understanding, 1503

Data Preparation, 1504

Modeling—and Types of Models, 1504

Testing/Evaluation, 1505

Deployment and Monitoring, 1505

Types of Machine Learning Models and Other Terminology, 1506

Supervised Learning, 1506

Unsupervised Learning, 1507

Reinforcement Learning, 1508

Artificial Neural Networks, 1508

Deep Learning, 1509

- Judicial Roles and High-Level AI Takeaways for Judges, 1511
 - There Are Many Different Methodologies and Designs, 1512
 - Most AI Is Iterative and Should Be Tested and Validated Continuously, 1513
 - Humans Are Always Involved, 1513
 - AI Predicts, It Does Not Conclude, 1515
 - Currently, Accuracy Often Depends on the Volume and Quality of Data, 1516
 - Some AI Is Not Explainable but More Transparent Methodologies Are Being Developed, 1517
 - The Heart of AI Is the Algorithm, 1517
 - Narrow AI Is Brittle, 1519
 - AI Is Also Nimble, 1519
 - Generative AI Can Hallucinate, 1519
 - AI Is Biased, 1520
- Bias, 1520
 - Forms of Algorithmic Bias, 1521
 - Statistical Bias, 1521
 - Moral Bias, 1522
 - Training Data Bias, 1522
 - Inappropriate Focus, 1523
 - Inappropriate Deployment, 1524
 - Interpretation Bias, 1524
 - Unwitting Human Bias, 1525
 - Intentional Bias, 1526
 - Overfitting and Outliers, 1527
 - Probing for Bias, 1528
- Algorithms That Predict Human Behavior, 1530
- Algorithms for Lawyers and Non-Lawyers:
 - Technology-Assisted Review and . . . Legal Practice?, 1536
- AI and Discovery, 1539
- Judges as AI Gatekeepers, 1542
 - Federal Rules of Evidence 401–403, 702, 901–902, 1542
 - Context, 1544
 - Case-Specific Reliability, 1545
 - Inapt Factors, 1545
 - Underlying Bias, 1546
- Daubert*, 1546
 - Testing, 1547
 - Peer Review, 1547
 - Error Rates, 1549
 - Standards Controlling an AI Application’s Operation and Maintenance, 1549

Reference Guide on Artificial Intelligence

Acceptance, 1549
Proprietary Algorithms, 1549
Deepfakes and Evidence, 1550
Conclusion, 1554
Glossary of Terms, 1555
Acknowledgments, 1560

Introduction

Artificial Intelligence (AI) now permeates our lives in seen and unseen ways. It informs how we shop, drive, recreate, and work. It guides supply chains and helps with medical diagnoses and documentation. Just as AI is transforming the economy, healthcare, and American society, it will also transform the practice of government and law. Lawyers use AI platforms for discovery, legal research, and drafting. At least seventy-five countries use facial recognition for domestic security and law enforcement purposes.¹ AI is used to determine travel patterns, link suspects with crime scenes, and populate watch lists. Between 2011 and 2019, the FBI used its facial recognition algorithm to search federal and state databases, including some state driver's license databases, over 390,000 times.² Modern militaries and intelligence services are integrating elements of AI into logistics, weapons, and information systems. "Generative AI" can create language loosely in the style of famous writers, simulate paintings reminiscent of those of the masters, imitate voices and generate fake video of real people, perform academic research at varying levels of sophistication and accuracy, and pass professional exams. The National Security Commission on Artificial Intelligence (NSCAI) has predicted that "[t]he development of AI will shape the future of power."³

Many of these uses of AI may find their way into U.S. courts. Some already have. Not only will courts see AI-generated evidence in civil and criminal cases, some state courts and jurisdictions already choose to use various algorithmic risk assessments in bail, parole, and even sentencing decisions.⁴ Judges may need to determine whether to admit AI outputs into evidence or to use AI in judicial decision making. They may also be called upon to adjudicate disputes related to how AI is used in the real world, such as the implications of the use of recommendation algorithms by social media platforms to direct content to particular users, as in the 2023 Supreme Court cases *Twitter v. Taamneh* and *Gonzalez v. Google*.⁵ To address these challenges, judges must understand how AI works, its

1. Steven Feldstein, The Global Expansion of AI Surveillance 1 (Carnegie Endowment for International Peace, Sept. 17, 2019), <https://perma.cc/C5H2-T2ZZ>; National Security Commission on Artificial Intelligence (NSCAI), Interim Report 12 (Nov. 2019), <https://perma.cc/RWP3-ZXP8>.

2. U.S. Gov't Accountability Office, GAO-19-579T, Face Recognition Technology: DOJ and FBI Have Taken Some Actions in Response to GAO Recommendations to Ensure Privacy and Accuracy, But Additional Work Remains (June 4, 2019).

3. NSCAI Interim Report, *supra* note 1, at 9.

4. "Liberty at Risk: Pre-Trial Risk Assessment Tools in the U.S.," Electronic Privacy Information Center 1 (updated Sept. 2020), <https://archive.epic.org/LibertyAtRiskReport.pdf>. As noted below, criminal risk assessments may not or do not all use machine learning but may increasingly do so in the future.

5. *Gonzalez v. Google*, LLC, 2023 U.S. LEXIS 2059 (May 18, 2023); *Twitter v. Taamneh*, 143 S. Ct. 762 (2023).

applications, its implications for the fact-finding process, and its risks. Judges should be able to answer the following four question-sets in context:

- How is the AI model being used (in the real world) or proposed to be used (in court or to inform judicial decisions)? Is that use the purpose for which the model was designed?
- Does the fact finder understand the AI's strengths, limitations, and risks, such as bias?
- Is the AI application authentic, relevant, reliable, and material to the issue at hand, and is its use or admission consistent with the Constitution, statutes, and the Federal Rules of Evidence (or state equivalents)?
- Has an AI algorithm, a human, or some combination of the two made any "judicial decision," and, in all cases, has that decision been documented in an appropriate and transparent manner allowing for judicial review and appeal?

This reference guide addresses these questions by providing technical background judges should possess to address legal issues that may arise surrounding AI-generated evidence, litigation over AI-enabled vehicles and applications, and proposed AI judicial tools.

However, describing AI is a challenge. The field incorporates many different technologies, functions, and uses. Almost any application or tool that uses computers, algorithms, and data might be described as AI or touch upon AI. AI is also a moving target. AI is by design iterative. This reference guide is intended as a foundation with which judges and court staff can more easily identify and assess this rapidly evolving field by presenting core concepts and terms. Specific applications and methodologies, however, will invariably change and change rapidly over time, including, for example, methods to better identify and understand AI-generated outputs (explainability) as well as validate digital imagery and voice patterns using AI-enabled tools. We do not provide legal judgments about the use of different AI applications. In discussing how AI is used today and may be used in the future, we do not endorse that use in any particular context or application—fact-intensive questions. Rather, we identify core technical concepts and legal issues, so that when judges must decide whether to admit AI applications into evidence or to use AI in a judicial determination, they decide wisely and fairly. Making these decisions requires judges and litigators to know enough about AI to ask the right questions, at the right moment, in the right depth.

We conclude this introduction with a caveat and a statement of purpose. AI is a field of technologies that is changing at an exponential pace. Today's cutting-edge application may be commonplace tomorrow, or obsolete. New methods will replace old methods. Therefore, it is not important for judges to understand how the IBM computer Deep Blue defeated world champion Gary Kasparov in chess

in 1997. It is important to know that a machine beat a human, which means that machines can be better than humans at certain tasks for which humans otherwise use human intelligence. Likewise, it is not important for judges to know how to label or code training, validating, or testing data. It is important to know that the quality and qualities of AI datasets can impact AI accuracy because datasets, like humans, possess bias. In a specific litigation context, expert testimony may be required on the subject. The goal of this reference guide is not to make judges scientific experts on AI or specific AI applications or methodologies; the field will change before this reference guide goes to print. Judges do not need to be coders or a technologist to ask good questions; they need to know good questions to ask. That is what judges do. We hope this reference guide provides a helpful background and sound framework for judges to do so.

Roadmap

This reference guide starts with a brief history and overview of the constellation of technologies associated with AI. The entry point to AI will vary for most jurors and judges; this section offers a quick level set. The next sections take a deeper look at the basics of machine learning, the underpinning of what today is generally known as AI. The reference guide then considers the roles judges will play regarding AI, including overseeing discovery, acting as evidentiary gatekeepers, interpreting law's application to AI contexts, translating complex questions of technology and law into case law, and considering whether to use AI tools in judicial decision making. We provide some high-level technological takeaways about AI for judges, then focus on two areas likely to generate adjudication: AI bias, and algorithms that predict human behavior. All of this technical knowledge is a predicate to judges understanding the myriad and complex ways AI will come before courts. Informed by this knowledge, the reference guide concludes by addressing how AI may change the practice of law as well as discovery and evidentiary considerations presented by AI, including AI deepfakes. Throughout, the reference guide identifies with bullet points threshold questions judges should ask about AI in context. They are threshold questions because each in turn should lead to additional contextual questions about specific AI applications. A glossary of terms follows the reference guide narrative.

An Overview of AI

What Is AI?

AI scholars and practitioners use multiple definitions of AI, a fluid and expansive field. The National Security Commission on AI provided one helpful definition:

Reference Guide on Artificial Intelligence

AI is not a single piece of hardware or software, but rather, a constellation of technologies that gives a computer system the ability to solve problems and to perform tasks that would otherwise require human intelligence.⁶

Each of the components constituting a particular AI system may be subject to legal challenge and validation. Moreover, because many AI components like algorithms are iterative, “learning” as they go, the field is constantly evolving and at a rapid pace. Case law will necessarily need to evolve and adapt as well.

A 2017 Belfer Center Study on AI and national security similarly emphasized that the field of AI is composed of multiple methodologies and technologies:

Artificial Intelligence (AI) is a science and a set of computational technologies that are inspired by—but typically operate quite differently from—the ways people use their nervous systems and bodies to sense, learn, reason, and take action.⁷

These definitions emphasize that artificial intelligence, while often analogized to human thinking with terms like “artificial neural network” and “neurons,” is achieved by mathematical computation: AI identifies, collects, aggregates, analyzes, and derives meaning from data, or performs tasks that rely on data. The AI is not “learning” as humans do; it is using mathematical models and other techniques to better perform or optimize tasks it has been programmed and trained to perform. AI learning is generally statistical in nature. However, as noted throughout this reference guide, developing and creating AI is a very human endeavor that reflects human choices and biases. Machines may be programmed to act autonomously, like a chatbot creating its own code and responses, but it is humans who write the source code that enables the bot to work and humans who select and define the data on which the bot is trained, tested, and validated.

Therefore, a single, capacious definition of AI is elusive, and rightly so. Depending on the AI application, a system will contain different technologies and methodologies, train on different data, and encounter different real-world data inputs in the field, all of which means each application will also have different strengths and weaknesses. At present, case law on AI is limited,⁸ but even as it develops, rather than simply relying on precedent, courts will still need to examine each AI system and its real-world application to specific facts and conditions. However, at least three unifying characteristics give machines “AI”: they

6. National Security Commission on Artificial Intelligence (NSCAI), Interim Report 8 (Nov. 2019), <https://perma.cc/RWP3-ZXP8>.

7. Greg Allen & Daniel Chan, Artificial Intelligence and National Security, Belfer Center for Science and International Affairs, Harvard Kennedy School (July 2017), <https://perma.cc/RRF9-29VQ>.

8. We have compiled an appendix list of illustrative cases in our longer publication, James Baker, Laurie Hobart, & Matthew Mittelsteadt, *An Introduction to Artificial Intelligence for Federal Judges*, (Federal Judicial Center 2023), <https://perma.cc/7DJT-T9D2>.

have (1) the capacity to “learn” (with the caveats above) and (2) the capacity to act (3) based on seed coding (algorithms) designed by humans.

A Brief History from Turing to Today: Waves of AI Development

Popular and scientific literature identifies several benchmark events in AI development. In 1950, the English computer scientist and Bletchley Park code breaker Alan Turing wrote an article, “Computing Machinery and Intelligence.” He asked, can machines think, and can they learn from experience as a child does? “The Turing Test” was Turing’s thesis advisor’s name for Turing’s experiment testing the capacity of a computer to think and act like a human. A computer would pass the test when it could communicate with a person in an adjacent room without the person realizing they were communicating with a computer.

In 1956, Dartmouth College hosted the first conference to study AI.⁹ The host, Professor John McCarthy, is credited by many with coining the term “Artificial Intelligence.”¹⁰ The funding proposal submitted to the Rockefeller Foundation stated,

We propose that a 2-month, 10-man study of artificial intelligence be carried out. . . . We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together.¹¹

Notwithstanding this optimistic start, progress in the field was neither linear nor exponential. It occurred in fits and starts. As a result, AI development went through a series of “AI winters,” periods of low funding and low results. In the past twenty years, however, AI has emerged as one of the transformative technologies of the twenty-first century.

Experts have described three waves of AI development. The first wave of AI machine learning consisted of if-then linear learning, a process that relies on the brute-force computational power of modern computers. With linear learning, a computer is in essence “trained” that if something occurs, then it should take a countervailing or corresponding step. This is how the IBM computer Deep Blue beat Gary Kasparov in chess in 1997, a significant AI milestone. The computer was optimizing its computational capacity to search through possible positions and use decision trees to weigh possible moves in response to Kasparov’s actual and

9. *Artificial Intelligence (AI) Coined at Dartmouth*, Dartmouth College, <https://perma.cc/856H-6NYD> (last visited Jan. 1, 2025).

10. *Id.*

11. Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* 6 (2014).

potential moves.¹² It did so with the knowledge of all of Kasparov’s prior games, while on the clock in real time. Deep Blue was an impressive demonstration of computational force, a near instantaneous series of if-this-then-that calculations.

We are in (at least) a second wave of machine learning now, the benchmark of which is AlphaGo, the Google computer that beat the world’s best Go player in 2016. The AlphaGo victory was a milestone not just because Go is a more complex, multidimensional game than chess but because AlphaGo won using reinforcement learning: it got better at the game by playing it. AlphaGo improved with experience, adjusting its own decisional weights internally—in the so-called “black box” of internal machine calculation—without training data or other if-this-then-that learning.¹³ Surpassing brute force computational power, this was a machine optimizing its capacity. Was it “thinking”? No. But was it “learning”? Yes.

What changed? Experts point to several factors working synergistically—specifically, the development of complex algorithms, strides in computational speed, the invention of new sensors, an explosion in data, and the advent of cloud computing and machine learning.

Complex Algorithms

Merriam-Webster’s Collegiate Dictionary defines an algorithm most “broadly” as “a step-by-step procedure for solving a problem or accomplishing some end.”¹⁴ A common example is a recipe for a meal. More narrowly, an algorithm is “a procedure for solving a mathematical problem . . . in a finite number of steps that frequently involves repetition of an operation.” In the AI context, algorithms are rigorously defined steps, written in software, to find, sort, and look for patterns in data. Algorithms are “the set of rules a machine (and especially a computer) follows to achieve a particular goal.”¹⁵

Today’s algorithms can be (but are not always) much more complex than simple recipes. The Google enterprise utilizes over two billion lines of code, while the search engine operates with 200,000–300,000 lines of code.¹⁶ The engine itself combines multiple tools and algorithms that web index (internet catalogue and map), embed potentially hundreds of search factors, and rank pages and references.

12. Adam Rogers, *What Deep Blue and AlphaGo Can Teach Us About Explainable AI*, Forbes (May 9, 2019), <https://perma.cc/2EJW-JNBB> (“The 1997 version was capable of searching between 100 and 200 million positions per second, depending on the type of position, as well as a depth of 20 or more pairs of moves.”).

13. David Silver et al., *Mastering the Game of Go Without Human Knowledge*, 550 *Nature* 354 (Oct. 19, 2017), <https://doi.org/10.1038/nature24270>.

14. Merriam-Webster’s Collegiate Dictionary, <https://perma.cc/82FG-EJRD> (last visited Aug. 18, 2022).

15. *Id.*

16. Rachel Potvin, *Why Google Stores Billions of Lines of Code in a Single Repository*, YouTube (Sept. 14, 2015), <https://perma.cc/5S56-5E7Z>.

By some accounts “the algorithm,” meaning the collective algorithms and tools, are adjusted 500–600 times a year.¹⁷ Thus, there is no single, final Google search algorithm. For many AI applications, not just Google, the algorithm one uses today will be different from the algorithm one uses tomorrow.

Computational Speed

The foundational component of a computer is the central processing unit (CPU), which today is composed of billions of electrical components called transistors. The more transistors in a CPU, the more data that can be processed in smaller packages with greater speed.¹⁸ Increased computational power is directly tied to the miniaturization of the transistor. So far, the number of transistors in a CPU has followed a trend called “Moore’s Law,” “the observation that the number of transistors on an integrated circuit will double every two years with minimal rise in cost.”¹⁹ By one estimate, a 1985 Cray Supercomputer, which took up sixteen square feet, would have had to be expanded to 80,000 square feet, or the size of an office building on two acres of land, to have the same processing speed as a 2020 iPhone 12.²⁰ Humans interact and program CPUs with programming languages that are ultimately turned into ones and zeros (the only thing a CPU understands) through different layers of software. One of the defining characteristics of AI is its capacity to perform tasks and process data at “machine speed.” While all machines might be said to operate at “machine speed,” this term is used by many to mean in effect “instantaneous” speed, distinguishing the capacity of machines to process information and make calculations from the capacity of humans performing the same or similar tasks.

Sensors

The development of sensor technology, such as that found in driverless cars, cell phones, and home devices, has resulted in more data and more applications for using that data to inform and influence human behavior. Personal assistants like Siri, Watson, and Alexa all use sensors to collect data.

17. See Ryan Shelley, *3 Things to Do After a Major Google Algorithm Update*, Search Engine Land (Oct. 18, 2016), <https://perma.cc/AQ25-SBZA>.

18. For example, in 1971 the Intel 4004 processor contained 2,300 transistors and by 2010 an Intel Core processor had over 560 million transistors, <https://perma.cc/TA5X-FQ5E>.

19. *Moore’s Law*, Intel Newsroom, <https://perma.cc/VP8Q-9RD6>.

20. Chandra Steele, *Space Wars: 80s Cray-2 Supercomputer vs. Modern-Day iPhone*, PC Magazine (Nov. 23, 2022), <https://perma.cc/5AGF-L9TE>.

Data

Data drives the AI revolution. For many though not all types of current AI modeling, the more data one has the easier it is to train a computer system to perform a task or solve a problem, and likely the more accurate the result.²¹ As discussed below, the quality of the data, as well as the metrics selected in designing algorithms, will also affect accuracy and the degree to which different forms of bias will affect accuracy.²²

Cloud Computing

The advent of cloud computing allows more data to be stored on a permanent and centralized, and thus retrievable, basis. As the Supreme Court encountered in *Carpenter v. United States*, where a criminal prosecution relied on historical cell phone location records that provided “a comprehensive chronicle of the user’s past movements,”²³ data can persist for years, even forever, if not purposefully deleted as a matter of law or policy. (The lifespan of data might also be limited by the expense of storing it and the possibility that future software or computer systems might not be able to read its format or encoding.)

Machine Learning

A critical technological advancement has been the development of machine learning, which refers to different methodologies to program software-driven machines to learn on their own and thus improve and optimize their functions. After the Dartmouth conference in 1956, the AI field “split into two camps: one focused on symbolic systems, problem solving, psychology, performance, and serial architectures, and the other focused on continuous systems, pattern recognition, neuroscience, learning, and parallel architectures.”²⁴ That second camp, or AI subfield, was machine learning. Much machine learning research is

21. Large language models, discussed below, rely on vast quantities of data, but experts are also developing models that do not depend on large quantities of data. For example, an AI “information lattice” model that helps humans without a musical background compose music was trained on a set of 370 Bach compositions. See Michael O’Boyle, *(Re)discovering Music Theory: AI Algorithm Learns the Rules of Musical Composition and Provides a Framework for Knowledge Discovery*, U. of Ill. Urbana-Champaign News (Jan. 27, 2023), <https://perma.cc/BA62-KK46>.

22. See Remarks of Nisheeth Vishnoi at Yale Cyber Leadership Forum, “Session #1: Big Data, Data Privacy, and AI Governance” (Feb. 18, 2022), <https://perma.cc/62SK-L3SW>.

23. 138 S. Ct. 2206 (2018).

24. Kush R. Varshney, *Trustworthy Machine Learning* 1 (2022), <https://perma.cc/LHL2-ANFP>. Dr. Varshney, an IBM Fellow, directs the Human-Centered Artificial Intelligence and

predicated on trying to mimic the human brain (literally, in the case of efforts to replicate the brain using 3D printers) or with neurological metaphors such as “artificial neural networks.” While AI may mimic, and in some cases outperform, human intelligence, it is not actual human intelligence. It is machine capacity and optimization, hence the preferable term: Human-Level Machine Intelligence (HLMI). Machine learning is explained in more detail, below.

Specialists describe the AI of today as “narrow,” generally good at addressing discrete tasks or “application areas[,] such as playing strategic games, language translation, self-driving vehicles, and image recognition.”²⁵ Narrow AI capabilities exploded in the early 2010s with a machine learning technique called “deep learning,”²⁶ described in greater depth below.

Beyond narrow AI, computer engineers contemplate the emergence of Artificial General Intelligence, or AGI. Artificial General Intelligence is an AI multitasking capacity that can serve multiple purposes. Much like a human, AGI would be able to understand and perform multiple tasks and shift from task to task as needed. Most analysts foresee AGI arriving, if it arrives at all, as a stage in development, like the advent of flight—not necessarily a moment in time, like the Soviet launch of the Sputnik satellite in 1957. AGI will present more complex legal questions than narrow AI. A system that can write and rewrite its own code as well as shift from task to task will be harder to regulate, requiring courts and legislators to wrestle with questions of accountability and responsibility for actions the AI takes or information it provides.

Indications of a potential third wave of AI machine learning were being discussed in 2016, the year of AlphaGo. As the *National Artificial Intelligence Research and Development Strategic Plan* reported in October of that year,

[t]he AI field is now in the beginning stages of a possible third wave, which focuses on explanatory and general AI technologies. . . . If successful, engineers could create systems that construct explanatory models for classes of real world phenomena, engage in natural communication with people, learn and reason as they encounter new tasks and situations, and solve novel problems by generalizing from past experience.²⁷

Trustworthy Machine Intelligence teams at IBM, and was the founding co-director of the IBM Science for Social Good initiative from 2015–2023.

25. Executive Office of the President, National Science and Technology Council, Committee on Technology, Preparing for the Future of Artificial Intelligence 7 (Oct. 2016), <https://perma.cc/BD4K-VC8G>.

26. *Large, Creative AI Models Will Transform Lives and Labour Markets*, The Economist (Apr. 22, 2023), <https://perma.cc/7YTH-R25H>.

27. National Science and Technology Council, Networking and Information Technology Research and Development Subcommittee, The National Artificial Intelligence Research and Development Strategic Plan 14 (Oct. 2016).

Today, we may well be on the rising swell of a third wave, and somewhere between narrow AI and AGI. In late 2022, OpenAI released ChatGPT, which stunned users with its ability to respond conversationally to prompts and to generate relatively sophisticated text and research products across subject matter, though not without error. ChatGPT is a large language processing model (LLM), a type of “generative AI.” Generative AI models use machine learning to generate or create new images, text, voices, videos, computer code, or other content; some are even being developed to discover new molecules.²⁸ Yet, as might be expected from models trained on text and images from the internet, ChatGPT and other generative AI models have been shown to create false narratives and reports, to embed racism, nationalism, and other biases, and even to generate fake legal cases. Even when trained on factually accurate data, generative AI models such as ChatGPT have produced false or counterfactual outputs, often referred to as “hallucinations.”

Big tech companies are working on AI products similar to ChatGPT,²⁹ with an emphasis on “big”: Thus far, the success of LLM models like ChatGPT has been contingent on vast amounts of training data (estimated at 45 terabytes for ChatGPT³⁰) as well as huge compute capacity and gigawatt-hours of electricity required to train and test the models on that data.³¹ All of these factors come at significant monetary expense as well. AI researchers are developing methods to reduce the size of generative models, for example by decreasing the number of mathematical connections between artificial neurons.³² Generative AI has been improving “at a dizzying pace, with new models, architectures, and innovations appearing almost daily.”³³

No matter how sophisticated and impressive, ChatGPT and other LLMs, with their errors and limitations, are not yet Artificial General Intelligence, or human level intelligence, though they approach it.³⁴ AGI contemplates that a computer linked to the internet, the Cloud, and the Internet of Things (IoT) might perform multiple tasks, solve problems, or answer questions generally, moving fluidly from one function to the next. Commentators ask three

28. *What Is Generative AI?*, IBM, <https://perma.cc/9J6Z-JHUP>.

29. *Large, Creative AI Models Will Transform Lives and Labour Markets*, The Economist, *supra* note 26.

30. *What Is Generative AI?*, McKinsey & Co. 4 (Apr. 2, 2024), <https://perma.cc/8RMA-PV3Z>.

31. *Large, Creative AI Models Will Transform Lives and Labour Markets*, The Economist, *supra* note 26; *What Is Generative AI?*, IBM, *supra* note 28.

32. *The Bigger-Is-Better Approach to AI Is Running out of Road*, The Economist (June 21, 2023), <https://perma.cc/Y22K-SCQZ>.

33. *What Is Generative AI?*, IBM, *supra* note 28.

34. The OpenAI website acknowledges this, stating “We believe our research will eventually lead to artificial general intelligence, a system that can solve human-level problems. Building safe and beneficial AGI is our mission.” <https://openai.com/research/overview> (last accessed Oct. 18, 2024).

threshold questions when it comes to general AI: Will it happen? When will it happen? And will it be a good or bad thing when it does?

From 2015 to 2016, a group of scholars associated with the Oxford Future of Humanity Institute, AI Impacts, and Yale University surveyed “all researchers who published at the 2015 NIPS and ICML [Workshop on Neural Information Processing Systems and International Conference on Machine Learning] conferences (two of the premier venues for peer-reviewed research in machine learning).”³⁵ The survey asked respondents to estimate when human-level machine intelligence (HLMI) would arrive. The study did not define AGI but stipulated that “Human-Level Machine Learning is achieved when unaided machines can accomplish every task better and more cheaply than human workers.” Three-hundred and fifty-two researchers responded, a return rate of 21%. The results ranged across the board: from fairly soon to never to beyond 100 years. What is noteworthy is that the “aggregate forecast gave a 50% chance of HLMI occurring within 45 years and a 10% chance of it occurring within 9 years.” The two countries with the most survey respondents were China and the United States. The median response for the Americans was seventy-six years, for the Chinese, twenty-eight.³⁶ As the survey indicates, experts do not agree on whether or when we will get to AGI.

Some experts believe that the powers and potential of AI are exaggerated. The authors of this reference guide do not share that view. Neither do the governments and companies dedicating billions of dollars to the development of AI. AI tools and methods will continue to change rapidly. The same factors that led to current AI, including data, computational capacity, and innovative modeling, will generate the next waves of AI, and those factors will only grow. The courts, like other elements of society, must adjust to AI, just as they previously adjusted to computers and electronic filing. Preparation starts with an understanding of what AI is, and is not, and the confidence that, explained in plain language, AI, its strengths, and its weaknesses, can be understood by judges, litigants, and jurors.

A Survey of AI Applications “Today”

Contemporary “narrow AI” can be defined as the “ability of computational machines to perform singular tasks at optimal levels, or near optimal levels, and

35. Katja Grace et al., Viewpoint: *When Will AI Exceed Human Performance? Evidence from AI Experts*, 62 J. of Artificial Intelligence Res. 729–54 (July 31, 2018), <https://doi.org/10.1613/jair.1.11222>.

36. *Id.*

usually better than, although sometimes just in different ways [than], humans.”³⁷ Under this umbrella come many single-purpose technologies, e.g., the AI that enables facial recognition and driverless vehicles. These technologies are “intelligent” in only one domain, limiting their ability to be used for multiple purposes or deal with certain complex situations.

Applications of machine learning today fall under three broad categories:

(1) cyber-physical systems, (2) decision sciences, and (3) data products. Cyber-physical systems are engineered systems that integrate computational algorithms and physical components, e.g. surgical robots, self-driving cars, and the smart grid. Decision sciences applications use machine learning to aid people in making important decisions and informing strategy, e.g. pretrial detention, medical treatment, and loan approval. Data products applications are the use of machine learning to automate informational products, e.g. web advertising placement and media recommendation.³⁸

Current AI is particularly good at correlating, connecting, and classifying data; recognizing patterns; and weighting probabilities—which is why it is good, and getting better, at tasks like facial recognition, image compression and identification, and voice recognition. But the field of AI is not static, it is moving at a seemingly exponential pace. That is why it is more important for judges to ask the right questions than to “master” a single iteration of AI or AI application. Binding and precedential case law will prove elusive.

At present, narrow AI can be “brittle,” by which engineers mean incapable of adapting to new circumstances on its own and lacking in situational awareness. To illustrate the point, AI philosophers like to point to the thought experiment known as the Trolley Problem, now more commonly conveyed as a crosswalk problem. In the problem, a driverless car loses its brakes at just the moment it is coming up to a crosswalk at speed. There are various pedestrians in the crosswalk of different ages and of different perceived virtues—for example, a pregnant woman and an armed robber carrying a bag of stolen money. The car’s computer must make a choice: swerve left, swerve right, brake, or drive ahead. An alert human driver, after first trying to brake, would in theory make a values-based, ethical choice about where to aim the car, likely at the presumed bank robber. The AI-driven car, on the other hand, unless it has been specifically trained to identify an “armed bank robber” or a “pregnant woman” and to adjust its decisional weights to favor one over the other in a choice scenario, will likely perceive the persons in the crosswalk as “persons in the crosswalk,” no more. Chances are the software code will select the path of least numerical damage. This weakness in current AI is especially important where an AI application is likely to

37. James E. Baker, *The Centaur’s Dilemma: National Security Law for the Coming AI Revolution* 34 (2020).

38. Varshney, *supra* note 24, at 2.

encounter changing or novel circumstances, like driving, or where there is an incentive for external actors to spoof or fool the AI application, as might be the case with military, intelligence, and law enforcement surveillance tools.

Many narrow AI applications are known to consumers who rely on them daily. If you shop on Amazon, you are using AI algorithms. Amazon back-propagates training data from purchases made on Amazon as well as data from individual consumers. Algorithms then identify patterns in the data and weight those patterns, allowing the algorithm to suggest (predict) additional purchases to the shopper. The algorithm adjusts as it goes based on the responses (or lack of responses) from recipients. This is an example of predictive big-data analytics. It is also an example of a push, predictive, or recommendation algorithm.

Why do companies use AI? Former Secretary of the Navy Richard J. Danzig explains:

... machines can record, analyze and accordingly anticipate our preferences, evaluate our opportunities, perform our work, etc. better than we do. With ten Facebook “likes” as inputs, an algorithm predicts a subject’s other preferences better than the average work colleague, with 70 likes better than a friend, with 150 likes better than a family member and with 300 likes better than a spouse.³⁹

Narrow AI is also embedded in mapping applications, which sort through route alternatives with constant, near-instantaneous calculations factoring speed, distance, and traffic to determine the optimum route from A to B. Then the application uses AI to convert numbered code into natural language telling the driver to turn left or right.

AI is also used to spot minute changes in stock pricing and to generate automatic sales and purchases of stock, as well as to spot anomalies that generate automatic sales and purchases. All of this is based on algorithms created and initiated by humans but programmed to act autonomously and automatically because the calculations are too large, the margins too small, and the speed too fast for humans to keep pace and make decisions in real time. Of course, as one trader’s algorithm gets faster, the next trader must either change their algorithm’s design, speed, or both to achieve an advantage, reducing the window of opportunity for real-time human control even further. AI machine learning and pattern recognition are also used for translation, logistics planning, and spam detection, among many, many more commercial applications. In 2017, Andrew Ng, the former chief scientist for Baidu, declared AI “the new electricity.”⁴⁰

39. See Richard Danzig, *An Irresistible Force Meets a Moveable Object: The Technology Tsunami and the Liberal World Order*, in *The World Turned Upside Down: Maintaining American Leadership in a Dangerous Age* (Nicholas Burns, Leah Bitounis, & Jonathon Price eds., 2017).

40. *Why AI Is the “New Electricity,”* Knowledge @ Wharton Staff (Nov. 7, 2017), <https://perma.cc/HD93-E9WH>.

A prominent illustration of emerging next-generation AI is the driverless vehicle, or more precisely, the software processes and hardware that enable driverless vehicles. AI empowers driverless cars by performing myriad data input and output tasks simultaneously, as a driver does, but in a different way. Human drivers rely on intuition, instinct, experience, and rules to drive—seemingly all at once—using the neural networks of the brain. In driverless cars, sensors instantaneously feed computers data based on speed, conditions, and images of the sort ordinarily processed by the driver's eyes and brain. The car's software processes the data to determine the best outcome based on probabilities and based on what it has been programmed to understand and decide. This requires constant algorithmic calculations that a human actor could not make in real time. Luckily, humans do not rely on math to drive cars. They exercise their judgment and intuition, which is why alert drivers generally handle situational change better than AI applications do. On the other hand, AI can possess 360-degree vision and does not fall asleep at the wheel, text while driving, or drive drunk. Driverless cars will and do make different kinds of mistakes than humans—but in theory and presumably in practice, fewer of them.

One of the most successful applications of AI to date is found in the field of medical diagnostics. Here, narrow AI's capacity to detect and match patterns and find anomalies has led to breakthroughs in the detection of tumors as well as the onset of diabetic retinopathy. In places like India, where there is a shortage of ophthalmologists, the use of such screening diagnostics can help prioritize access to doctors and treatment by identifying at-risk patients.⁴¹ Studies indicate that AI can be or may soon be more accurate than humans in detecting certain cancerous tumors or making prognostic findings about their growth and treatment.⁴² However, that is not the same as saying that humans are prepared to rely on AI alone, or wish to receive medical diagnoses from machines rather than doctors.

Law enforcement authorities use AI for predictive policing and surveillance. Some seventy-five countries use AI-powered surveillance, including many liberal democracies.⁴³ According to a 2019 Government Accountability Office report, the FBI has logged hundreds of thousands of searches of its facial recognition system, which has access to 641 million face photos.⁴⁴ The FBI

41. Cade Metz, *India Fights Diabetic Blindness with Help from AI*, N.Y. Times, Mar. 10, 2019, <https://www.nytimes.com/2019/03/10/technology/artificial-intelligence-eye-hospital-india.html>.

42. See Elizabeth Syoboda, *Artificial Intelligence Is Improving the Detection of Lung Cancer*, 587 Nature S20 (Nov. 18, 2020), <https://doi.org/10.1038/d41586-020-03157-9>; Nadia Jaber, *Can Artificial Intelligence Help See Cancer in New, and Better Ways?*, National Cancer Institute (Mar. 22, 2022), <https://perma.cc/A8HQ-5PW8>.

43. Steven Feldstein, *The Global Expansion of AI Surveillance* 1-2 (Carnegie Endowment for International Peace, Sept. 17, 2019), <https://perma.cc/C5H2-T2ZZ>; National Security Commission on Artificial Intelligence, Interim Report 12 (Nov. 2019), <https://perma.cc/RWP3-ZXP8>.

44. Drew Harwell, *FBI, ICE Find State Driver's License Photos Are a Gold Mine for Facial-Recognition Searches*, Wash. Post, July 7, 2019, <https://perma.cc/3WSW-MRSG>. See U.S. Gov't

reported its system has proven 86% accurate at finding the right person, if a search was able to generate a list of fifty possible matches.⁴⁵

Generative AI differs from traditional predictive models,⁴⁶ such as those used in driverless cars, shopping algorithms, and image recognition. Those predictive models classify data, identify patterns, and use statistics to predict or match other real-world examples with the data provided: For example, facial recognition applications are trained to identify images of humans and predict which images from a data bank might most closely match a real-world photo of a person.⁴⁷

Generative AI, while still using statistics and prediction,⁴⁸ creates new content. Early models developed in the 2010s could generate deepfake videos of political figures and excellent simulations of famous paintings.⁴⁹ ChatGPT has been used to author academic papers, poems, essays, books, and code;⁵⁰ GPT-4, released in March 2023, passed the July 2022 uniform bar exam with a score “approaching the 90th percentile” of human test-takers.⁵¹ ChatGPT and similar large language models (LLMs) use statistics to auto-fill or complete sentences. Today’s generative models train on datasets of sample text, images, or other content. The models “compress a dataset into a dense representation, arranging similar data points closer together in an abstract space.”⁵² For LLMs, the dataset is vast quantities of sample text, where common parts of words are represented in numbers called “embeddings.”⁵³ The LLM can “sample from this space to create something new while preserving the dataset’s most important features.”⁵⁴ An “attention” feature helps the LLM predict which embeddings from the training data will best match and complete a phrase and eventually a sentence.⁵⁵ Modern LLMs can predict the parts of a sentence in parallel or simultaneously.⁵⁶

Accountability Off., GAO-16-267, FACE Recognition Technology: The FBI Should Better Ensure Privacy and Accuracy (May 16, 2016); U.S. Gov’t Accountability Off., GAO-19-579T, Face Recognition Technology: DOJ and FBI Have Taken Some Actions in Response to GAO Recommendations to Ensure Privacy and Accuracy, But Additional Work Remains (June 4, 2019).

45. Harwell, *supra* note 44.

46. *What Is Generative AI?*, McKinsey & Co., *supra* note 30.

47. *Id.*

48. *Large, Creative AI Models Will Transform Lives and Labour Markets*, The Economist, *supra* note 26.

49. *See What Is Generative AI?*, IBM, *supra* note 28.

50. *Id.*; *Large, Creative AI Models Will Transform Lives and Labour Markets*, The Economist, *supra* note 26.

51. Debra Cassens Weiss, *Latest Version of ChatGPT Aces Bar Exam with Score Nearing 90th Percentile*, ABA Journal (Mar. 16, 2023), <https://perma.cc/7DDQ-QVCY>.

52. *What Is Generative AI?*, IBM, *supra* note 28.

53. *Large, Creative AI Models Will Transform Lives and Labour Markets*, The Economist, *supra* note 26.

54. *What Is Generative AI?*, IBM, *supra* note 28.

55. *Large, Creative AI Models Will Transform Lives and Labour Markets*, The Economist, *supra* note 26; *What Is Generative AI?*, IBM, *supra* note 28.

56. *Large, Creative AI Models Will Transform Lives and Labour Markets*, The Economist, *supra* note 26; *What Is Generative AI?*, IBM, *supra* note 28.

Reference Guide on Artificial Intelligence

Because they are based on statistical prediction and the content of training datasets, ChatGPT and similar models still make errors. While GPT-4 outperformed most bar takers, it still only answered 75.7% of the multiple-choice questions correctly and received scores around 4.2 out of 6 on the essays.⁵⁷ As noted earlier, ChatGPT has been shown to generate false information and research (sometimes referred to as “hallucinations”⁵⁸), and sometimes quite persuasively. Some federal judges have ordered that litigants disclose whether they have used generative AI in their research and briefing, and if so, affirm that they have verified its accuracy.⁵⁹ At least one state bar has cautioned members that the use of generative AI for legal research, depending on how it is used, may violate client confidentiality were the AI to incorporate the research into its ongoing dataset.⁶⁰ We surmise the advent of disputes over privilege waiver. The use of AI chatbots as legal “receptionists” may result in the creation of an attorney-client relationship, or appearance of one, without the attorney being aware.⁶¹

Generative AI can reproduce racism, nationalism, sexism, and every other bias found on the internet.⁶² The 2016 chatbot Tay, for example, generated and spread hate within hours of its public release, adopting the language and values found on Twitter.⁶³ (Similarly, non-generative social media “recommendation” or “push” algorithms that promote pre-existing content to certain users allegedly have been used to spread and amplify ethnic hatred, such as against the Rohingya in Myanmar;⁶⁴ disinformation, such as reported by Nobel Peace Prize recipient Maria Ressa in the Philippines⁶⁵ and widely observed in U.S. elections; and terrorist content, as the Supreme Court heard in *Gonzalez v. Google* and *Twitter v. Taamneh*.⁶⁶)

57. Weiss, *supra* note 51.

58. Karen Weise & Cade Metz, *When A.I. Chatbots Hallucinate*, N.Y. Times, updated May 9, 2023, <https://www.nytimes.com/2023/05/01/business/ai-chatbots-hallucination.html>.

59. See, e.g., Judge Brantley Starr, Mandatory Certification Regarding Generative Artificial Intelligence (N.D. Tex.), <https://perma.cc/GQ34-JLUM>.

60. Florida Bar Ethics Opinion 24-1 (Jan. 19, 2024), <https://perma.cc/PN5Z-E76Z>.

61. *Id.*

62. Marie Lamensch, Generative AI Tools Are Perpetuating Harmful Gender Stereotypes, Center for International Governance Innovation (June 14, 2023), <https://perma.cc/A56E-ZJKX>; Zachary Small, *Black Artists Say A.I. Shows Bias, With Algorithms Erasing Their History*, N.Y. Times (July 4, 2023), <https://www.nytimes.com/2023/07/04/arts/design/black-artists-bias-ai.html>.

63. Oscar Schwartz, *In 2016, Microsoft’s Racist Chatbot Revealed the Dangers of Online Conversation: The Bot Learned Language from People on Twitter—But It Also Learned Values*, IEEE Spectrum (Nov. 25, 2019), <https://perma.cc/5N4S-M8N9>.

64. Myanmar: Facebook’s Systems Promoted Violence Against Rohingya; Meta Owes Reparations, Amnesty International (Sept. 29, 2022), <https://perma.cc/53WW-3E3S>.

65. Maria Ressa, *Facts*, The Nobel Prize (2021), <https://perma.cc/CM45-Q4PH>.

66. *Gonzalez v. Google, LLC*, 2023 U.S. LEXIS 2059 (May 18, 2023); *Twitter v. Taamneh*, 143 S. Ct. 762 (2023).

The content created by generative AI models can be so realistic as to be misleading, raising significant concerns about models and their products being used wittingly and unwittingly for disinformation and misinformation, potentially polluting the internet at overwhelming levels. There is also significant concern that generative AI will be used to develop harmful and very real products, such as bioagents.⁶⁷

This survey of today's AI suggests that AI will both improve and necessarily be imperfect. For each potential application, humans must decide whether it is wise and fair to use AI in the first instance. Where it is employed, AI will generally be best used to augment rather than supplant human judgment. One issue AI policy makers, designers, and ethicists must resolve in context is how to structure human-machine teaming to allocate responsibility and accountability. Judges in turn will have to determine whether as a matter of law, or a matter of law and fact, the humans made the correct decisions: that is, whether the humans appropriately applied AI to a problem, selected an appropriate algorithm, and interpreted the results correctly. Judges will also have to consider to what extent, if any, they should rely on AI applications to inform their own decisions.

What Is Machine Learning?

Machine learning (ML) is a subfield⁶⁸ of and “one of the most important technical approaches to AI.”⁶⁹ As the name implies, it is a set of methods computers use to “learn,” both to perform tasks and to improve on that performance, by making predictions or classifications based on input data.⁷⁰ Dr. Kush Varshney writes:

The term machine learning was popularized in Arthur Samuel's description of his computer system that could play checkers, not because it was explicitly programmed to do so, but because it learned from the experiences of previous games. In general, machine learning is the study of algorithms that take data and information from observations and interactions as input and generalize from specific inputs to exhibit traits of human thought. Generalization is a

67. *How Generative Models Could Go Wrong: A Big Problem Is That They Are Black Boxes*, The Economist (Apr. 19, 2023), <https://perma.cc/7K7W-ZPEP>.

68. *What Is Artificial Intelligence (AI)?*, IBM, <https://perma.cc/L9YW-QN7V> (Machine learning and deep learning are subcategories of AI and “disciplines . . . comprised of AI algorithms which seek to create expert systems which make predictions or classifications based on input data.”).

69. Executive Office of the President, *supra* note 25, at 8.

70. Christopher Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable* (2d ed. 2022), <https://perma.cc/L98E-J4PM>; *What Is Artificial Intelligence (AI)?*, IBM, *supra* note 68.

process by which specific examples are abstracted to more encompassing concepts or decision rules.⁷¹

For example, the constellation of technologies that make up AI allow mathematically driven algorithms to perceive, break down, and connect patterns in data or digital sequences in images. Machine learning allows these machines to not only recognize patterns, such as those employed for facial recognition, shopping, and medical diagnosis, but to then on their own classify, or generalize, by associating certain patterns with desired outputs, e.g., potential matches in a facial recognition database.

How Machines Are Trained to Learn: The Machine Learning Life Cycle

Varshney outlines six stages in the development and life cycle of a machine learning system: (1) problem specification; (2) data understanding; (3) data preparation; (4) modeling; (5) evaluation; and (6) deployment and monitoring.⁷² As we will emphasize, humans are involved at every point along the continuum of that life cycle, infusing their choices into eventual AI outputs.⁷³ Experts also point to the importance of involving all stakeholders—i.e., those individuals or groups who might use or be affected by the AI application, including traditionally marginalized groups—at all stages of the machine learning life cycle.⁷⁴

71. Varshney, *supra* note 24, at 1. *See also* Executive Office of the President, *supra* note 25, at 8 (“Modern machine learning is a statistical process that starts with a body of data and tries to derive a rule or procedure that explains the data or can predict future data.”).

72. Varshney, *supra* note 24, at 14–22 (citing the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology); *see also* Nick Hotz, What is the Data Science Process?, <https://perma.cc/W8DU-SQ94> (identifying the CRISP-DM and other life cycle models for data science).

73. David Leslie, Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector, The Alan Turing Institute 16 (2019), <https://doi.org/10.5281/zenodo.3240529>.

74. Varshney, *supra* note 24, at 15, 17 (citing Meg Young, Lassana Magassa, & Batya Friedman, *Toward Inclusive Tech Policy Design: A Method for Underrepresented Voices to Strengthen Tech Policy Documents*, 21 Ethics & Information Tech. 89–103 (June 2019)), <https://doi.org/10.1007/s10676-019-09497-z>; Leslie, *supra* note 73, at 16; *see also* *Artificial Intelligence Ethics Framework for the Intelligence Community*, Version 1.0 (June 2020), <https://perma.cc/T6UT-JJL3> (“Identifying and addressing risk is best achieved by involving appropriate stakeholders. As such, consumers, technologists, developers, mission personnel, risk management professionals, civil liberties and privacy officers, and legal counsel should utilize this framework collaboratively, each leveraging their respective experiences, perspectives, and professional skills.”).

Problem Specification

The machine learning process begins with a question, problem, or task that a machine might address. The goal might be to automate a task or aid and improve upon human decision making,⁷⁵ such as detecting tumors or winnowing job applications. Even in this initial framing, humans are making choices that may reflect their own values and biases. Judges and lawyers will recognize this as analogous to the framing of a legal issue or question, or of a question in a survey or cross-examination. (Indeed, legal minds may find many connections and disjunctions between the rules and decision trees created by AI systems and those established by law.) Within the category of facial recognition algorithms, for example, it is a much different ethical undertaking to code and train a system to track a minority group, such as Uyghurs in China, than it is to develop a system to recognize missing children for Amber Alert purposes. The United States Intelligence Community has adopted an “Artificial Intelligence Ethics Framework” with lists of questions for government actors, including: “What is the goal you are trying to achieve by creating this AI? . . . Is there a need to use AI to achieve this goal? Can you use other non-AI related methods to achieve this goal with lower risk?” as well as “What benefits and risks, including risks to civil liberties and privacy, might exist when this AI is in use? Who will benefit? Who or what will be at risk? What is the scale of each and likelihood of the risks? How can those risks be minimized and the remaining risks adequately mitigated? Do the likely negative impacts outweigh likely positive impacts?”⁷⁶ The European Union has taken the approach of identifying certain AI uses as prohibited, and identifying other AI applications as “high risk,” requiring independent verification and testing.⁷⁷

At this initial stage, the human “problem owners” and AI technologists must also determine the metrics by which they will measure the success of the AI, another value-laden decision point.⁷⁸ For example, if the State Department decided to make foreign aid allocations to nation-states with the help of an algorithm, it might choose to measure success by how much a recipient state’s total GDP rose in future years or, instead, by what percentage of children received secondary-level education and certain vaccines. The two metrics might align, but they might not, or they might align but only to a point. Each metric also reflects different micro- and macro-level values.

75. Varshney, *supra* note 24, at 16.

76. *Artificial Intelligence Ethics Framework for the Intelligence Community*, *supra* note 74.

77. European Commission, *Regulatory Framework Proposal on Artificial Intelligence*, Sept. 29, 2022, <https://perma.cc/EUJ5-SJZT>.

78. Varshney, *supra* note 24, at 17; *see also* remarks of Nisheeth Vishnoi, Yale Cyber Leadership Forum, “Session #1: Big Data, Data Privacy, and AI Governance,” Feb. 18, 2022.

Data Understanding

The next step is selecting and collecting the relevant datasets that the machine learning model will use.⁷⁹ The data might be “numbers, photos, or text, like bank transactions, pictures of people or even bakery items, repair records, time series data from sensors, or sales reports.”⁸⁰ Some of the data will be “used as training data, or the information the machine learning model will be trained on.”⁸¹

When designing machine learning algorithms, engineers divide data into three sets: training data, validation data, and testing data.⁸² Training data is intended for the AI, curated and sometimes labeled so that the AI can analyze it, learn from it, and adjust its coding to ultimately form better predictions. Validation data is intended for the developer, who uses this data to stress test the model’s training and decide whether to update its settings and sensitivities. Testing data provides a final round of analysis; this data is used on the trained, tuned model to evaluate the model’s fit to the data and overall accuracy.

The selection of each of these sets of data is critical to the accuracy and fairness of the final product, hence the phrase “garbage in, garbage out.”⁸³ “Algorithmic bias” is discussed at length below, but by way of example here, two common types of bias in machine learning are “temporal bias” and “population bias.” Temporal bias concerns the timing of data collection and how that timing might affect accuracy.⁸⁴ This might occur if a university used an admissions AI algorithm that was trained on admissions data from the 1960s, before many American colleges accepted women. An algorithm so trained might not see fit to admit women (or rather, its outputs would not suggest doing so). Population bias occurs where some inputs are underrepresented and others overrepresented,⁸⁵ illustrated by underperforming facial recognition algorithms. If the training datasets do not contain sufficient numbers of photographs of faces of different skin colors and genders, there will be output disparity across “race” and gender. Social bias in algorithms can also “stem[] from prejudice: labels from historical human decisions contain systemic differences across groups.”⁸⁶ As discussed below, this concern comes up in criticisms of predictive policing algorithms and

79. Varshney, *supra* note 24, at 18–19; see also *CRISP-DM Help Overview*, IBM, <https://perma.cc/XB3C-N6TH>.

80. Sara Brown, *Machine Learning, Explained*, MIT Management Sloan School (Apr. 21, 2021), <https://perma.cc/R74J-SDDC>.

81. *Id.*

82. Tarang Shah, *About Train, Validation and Test Sets in Machine Learning*, Medium (Dec. 6, 2017), <https://perma.cc/6VRX-MUSP>.

83. Coined by Wilf Hey, computer scientist at IBM (Varshney, *supra* note 24, at 40).

84. Varshney, *supra* note 24, at 18.

85. *Id.*

86. *Id.* Quotations are used around “race” to emphasize the prevailing scientific and sociological understanding that race is a social construct, rather than a biological one.

criminal-risk assessment tools. Datasets might also include proxies, or pieces of data that might correlate with another type of information, such as a zip code for demographic information or, in some cultures, last names for religion.⁸⁷ Housing and employment data used in criminal-risk assessments have proved to be strong proxies for “race” and class.⁸⁸ An algorithm may not explicitly be programmed to make such a connection but it may nonetheless learn to do so. Proxies can be subtle (and sometimes unexpected by human programmers). For example, an individual’s online presence might serve as a proxy for age: An AI screening tool might distinguish between applicants with social media accounts and those without, or between types of social media accounts, both of which distinctions often reflect a user’s age.

Data Preparation

The data preparation stage of the machine learning life cycle involves multiple steps. Data scientists and data engineers first integrate data from multiple, disparate sets into one set in a single format; they then “clean” it by, among other things, filling in missing information, discarding other information, looking for outliers, and dropping features that might lead to data leakage (where the same data used to train the AI are encountered in actual use, potentially tainting the result by making it “easier” or more likely for the AI to find a match that would ordinarily occur in real-world conditions) or that might be unethical or illegal to consider.⁸⁹ The last step is feature engineering, or “mathematically transforming features to derive new features.”⁹⁰ In short, there is much room for “creativity” and choice⁹¹ at the data preparation phase, so the expertise, care, and worldview of the data scientist and data engineer have real impact.

Modeling—and Types of Models

Once the data are curated and prepped, engineers “choose a machine learning model to use, supply the data, and let the computer model train itself to find patterns or make predictions. Over time the human programmer can also

87. *Id.*

88. Chelsea Barabas et al., *An Open Letter to the Members of the Massachusetts Legislature Regarding the Adoption of Actuarial Risk Assessment Tools in the Criminal Justice System*, Berkman Klein Center for Internet & Society 3 (Nov. 9, 2017), <https://perma.cc/M77Z-MW7J> (citing Sonja B. Starr, *Evidence-Based Sentencing and the Scientific Rationalization of Discrimination*, 66 Stan. L. Rev. 803 (2014)).

89. Varshney, *supra* note 24, at 18–19.

90. *Id.* at 19.

91. *Id.*

tweak the model, including changing its parameters, to help push it toward more accurate results.”⁹²

In general, there are, at least currently, three categories of machine learning models: supervised learning, unsupervised learning, and reinforcement learning. Each category is described in Types of Machine Learning Models and Other Terminology, below.

Testing/Evaluation

After a model is made and the machine is trained on training data, it is next “stress-tested”⁹³ with new data. “Some data is held out from the training data to be used as evaluation data, which tests how accurate the machine learning model is when it is shown new data. The result is a model that can be used in the future with different sets of data.”⁹⁴ Experts are quick to point out that “leakage” between the training dataset and the testing dataset will undermine the value of the tests; if the machine is being shown the same data on which it was trained, its performance will seem better than it might be against new data. The remedy is clear: separate and distinct datasets; however, data lineage is not always known, clear, or available in the ways lawyers and judges understand chain of custody.

Deployment and Monitoring

The accuracy or reliability of an AI system can degrade over time if inputs shift away from training data⁹⁵ or immediately if new inputs and training data simply differ. As such, AI systems must be constantly and continuously monitored and tested even after they are deployed in the field.⁹⁶ A system trained to recognize video footage of battlefield tanks at night might, for example, misidentify a dark box that is actually an air-conditioner unit on the side of a building as a tank.⁹⁷ This is an example of brittleness and technical bias.

Precisely because some ML-driven AI continues to learn or degrade as it operates, the relevance and reliability of AI output can be moving targets. It is important to keep in mind that AI models might reinforce inaccurate or biased

92. Brown, *supra* note 80.

93. Varshney, *supra* note 24, at 22.

94. Brown, *supra* note 80.

95. Varshney, *supra* note 24, at 22.

96. James E. Baker et al., National Security Law and the Coming AI Revolution, Observations from a Symposium Hosted by Syracuse University Institute for Security Policy and Law and Georgetown Center for Security and Emerging Technology, Oct. 29, 2020 (2021), <https://perma.cc/8QKL-7W26>.

97. *Id.*

results as they learn. Opening discovery or testimony to AI/ML potentially opens the door to an array of subordinate questions involving methodologies, data, and testing. Judges may thus have to determine where it is essential for fact finders to understand the underlying methodologies, or just the conclusions or results derived from them.

In sum, not only does the AI field of study continue to evolve, but many individual AI applications also learn and change (for better or worse) over the course of their life cycles. Courts may need to test the reliability and validity of a given individual application both at precise points in its life cycle and against current best practices in the AI field. A court might examine not just how a particular AI application performed during its testing phase, but how it performed later in its life cycle, in the given, factually contested moment. A court might also consider whether a particular AI application meets the current best practices or standards in the AI field, such as those regarding the explainability and transparency of its computations and results. The evolving nature of AI also means that case law precedent may have limited or diminishing value when it comes to certain AI applications. Unlike polymerase chain reaction (PCR) DNA testing, for example, evidence generated by the same AI tool may vary in accuracy and reliability at different times.

Types of Machine Learning Models and Other Terminology

As alluded to above, there are three categories of machine learning models: supervised learning, unsupervised learning, and reinforcement learning. (Those categories, however, “simplify the enormous amount of complexity and variation in these systems.”⁹⁸)

Supervised Learning

“In supervised learning, the input data includes observations and labels; the labels represent some sort of true outcome or common human practice in reacting to the observation.”⁹⁹ For example, a dataset of photographs of lions

98. Ben Buchanan & Taylor Miller, *Machine Learning for Policymakers: What It Is and Why It Matters* 5 (Cyber Security Project, Belfer Center, June 26, 2017), <https://perma.cc/TZ4Q-UAX8>.

99. Varshney, *supra* note 24, at 1. See also Matthew Mittelsteadt, *Artificial Intelligence: An Introduction for Policymakers*, Mercatus Special Study (Mercatus Center at George Mason University, Feb. 2023).

and tigers might have photos that humans have labeled lion or tiger.¹⁰⁰ Running through its algorithms, the machine would learn on its own ways to discern whether new photos of big cats were lions or tigers (or perhaps neither). Here, “true outcome” should be read as referring to an objective reality. In this example, the “true outcome” of lion or tiger is straightforward. But one could imagine ways in which the very act of labeling a piece of data might be value-laden—for example, whether a particular applicant was a “strong” or “weak” candidate for a loan. “For the machine learning system to make even better decisions than people, true outcomes rather than decisions should ideally be the labels, e.g. whether an applicant defaulted on their loan in the future rather than the approval decision.”¹⁰¹

Unsupervised Learning

“In unsupervised learning, the input data includes only observations.”¹⁰² Algorithms “analyze and cluster unlabeled datasets” to “discover hidden patterns or data groupings without the need for human intervention.”¹⁰³ Unsupervised learning is a technique for teaching a computer to find links and patterns in large volumes of unlabeled data without a determined outcome in mind. In contrast, supervised learning matches a data point, such as an image, to a known database of labeled data. Unsupervised learning is often used for “exploratory analysis,” with “no simple goal” in mind.¹⁰⁴ It is harder to check or validate because there is no given answer or labeled dataset against which to check the model’s outputs.¹⁰⁵ The government might use an unsupervised learning methodology to search for meaningful patterns and hidden links in phone call records, travel patterns, or trade and commerce records indicating sanctions violations. Here, the algorithm is not searching for a particular number or face but for meaning in otherwise unstructured data. Importantly, it might also find connections without meaning, for example, by “matching” faces in a facial recognition application with similar backdrops or lighting. When this occurs within the program’s internal calculations or neural network as explained below, it may be difficult, or impossible, to discern that the “match” is based on a factor irrelevant to the output objective.

100. See Brown, *supra* note 80.

101. Varshney, *supra* note 24, at 18.

102. *Id.* at 1.

103. *What Is Unsupervised Learning?*, IBM (2020), <https://perma.cc/6D9A-2BWD>.

104. Gareth James et al., *An Introduction to Statistical Learning: with Applications in R* 373–74 (2017), https://doi.org/10.1007/978-1-4614-7138-7_1.

105. *Id.*

Reinforcement Learning

“In reinforcement learning, the inputs are interactions with the real world and rewards accrued through those actions rather than a fixed dataset.”¹⁰⁶ Reinforcement learning introduces a “machine learning agent,” either an incentive or a desired goal, into the algorithmic code that might cause the machine to weight or improve its outcome on its own, as in the case of AlphaGo.¹⁰⁷ A shopping algorithm, for example, might do this by automatically adjusting its code based on whether a recommendation is accepted, rejected, or ignored.

Artificial Neural Networks

One currently popular type of model that crosses all three major categories of machine learning (supervised, unsupervised, and reinforcement) is the artificial neural network. Artificial neural networks (ANN) mimic, or seek to mimic, the brain through a set of algorithms.¹⁰⁸ Metaphorical “neurons” each mark a discrete point in the network where the system interprets a data input and responds to it, based on formulae programmed by humans, or, as the machine learns, by the machine.

Every neural network consists of layers of nodes, or artificial neurons—an input layer, one or more hidden layers, and an output layer. Each node connects to others, and has its own associated weight and threshold. If the output of any individual node is above the specified threshold value, that node is activated, sending data to the next layer of the network. Otherwise, no data is passed along to the next layer of the network. . . . Neural networks rely on training data to learn and improve their accuracy over time.¹⁰⁹

An oversimplified example drawn from IBM’s website¹¹⁰ illustrates the process: Sam wants to go surfing as often as possible but only in good conditions. Sam develops a simple algorithm to help him decide each morning whether it is a good day to go. One input will be the water temperature; one the height of the waves; another the weather conditions; and the last the number of recent shark sightings. Each of these inputs will be assigned a weight, or a multiplier, representative of how much Sam values it. The shark multiplier will be negative. If the sum of the inputs and their multipliers is greater than a predetermined value, that neuron will “fire” or turn “on,” and will send an electronic signal or output

106. Varshney, *supra* note 24, at 1.

107. See Buchanan & Miller, *supra* note 98, at 11–12.

108. *AI vs. Machine Learning vs. Deep Learning vs. Neural Networks: What’s the Difference?*, IBM, <https://perma.cc/9YRM-HKYB>.

109. *What Is a Neural Network?*, IBM, <https://perma.cc/JXN3-HF2A>.

110. See *id.* The website presents a similar example with mathematical formulas.

to the next neuron to start calculating whatever decision it helps the system make (whether Sam should wear a wetsuit, perhaps). If the sum is less than the predetermined value for “getting to surf (yes),” the neuron does not fire, and the output signal never reaches the next neuron.

Deep Learning

Neurons are metaphorically assembled in layers, that is, time-ordered sequences. If, in between the input and output layers, there are three or more (potentially many more) hidden layers, the ANN is called a “deep learning” model.¹¹¹ Deep learning can be used in supervised, unsupervised, or reinforcement learning models.

Neural networks are sometimes referred to as the “black box” of machine learning. Because the machine is learning and adjusting as it goes, engineers may not be able to determine with certainty what the machine is weighing, how, and with how many input and output layers within the internal neural network. This is important where, for example, coding or data bias may impact the quality of the ultimate predictive outcome. A theoretical¹¹² algorithm designed to predict recidivism, for example, might use many different data inputs as well as patterns it derived from the data of the general population in generating a predictive percentage risk that an individual defendant will return to prison. But with deep learning the judge and lawyers might not know what decisional weight, if any, was assigned to “race,” location, past convictions, the letter “S” in a defendant’s name, or population trends inapt to a specific defendant.

Although deep learning and neural networks are the most common AI methods, they are not the only ones. AI is a dynamic field, and new developments could change how algorithms process and ascribe meaning to data.

Importantly, increasingly there are methodologies that engineers can incorporate to make such internal calculations more transparent, or fully transparent, to users. Research is underway in academia and by organizations such as the National Institute of Standards and Technology (NIST) to enable more transparent neural networks that will explain AI outputs on a global or per-decision basis after the fact so that AI users and those who are impacted by AI outputs might more fully understand the parameters and weights that are or were applied within.¹¹³ There are different approaches for doing so. Decision-tree models, for

111. *What Is Artificial Intelligence (AI)?*, IBM, *supra* note 68. See also *What Is Machine Learning (ML)?*, IBM, <https://perma.cc/JSA9-CXM6>.

112. Current recidivism algorithms are often proprietary but may lack sophistication and complexity.

113. *AI Fundamental Research—Explainability*, NIST (June 16, 2022), <https://perma.cc/3YX5-JC8Q>.

example, might run through a series of “if-this, then-that” analyses to essentially map out the input-to-output run of the network.¹¹⁴ Counterfactual algorithms seek to identify factors that were determinative by showing which factors would change the output, e.g., by identifying factors that caused a neural gate to open or close.¹¹⁵ Depending on the algorithm and its use, the factors could be benign, like a color or a shape, or potentially problematic, like “race,” or a proxy for “race.” Generally, the more detailed and layered an algorithm, the more accurate it is likely to be in its outputs (assuming good underlying data and coding methodology), but also the more difficult it will be to explain the output. In addition, not every AI application warrants, or would seem to require as a matter of law, the same measure of explainability. What users and courts require or expect to understand about an AI application that generates a credit score, a thunderstorm warning, or a bail risk decision will likely be different.

Moreover, what is transparent may not be explainable if it comes in the form of mathematical modeling and equations. A 2021 NIST report notes “that the failure to articulate the rationale for an answer can affect the level of trust users will grant that system. Suspicions that the system is biased or unfair can raise concerns about harm to oneself and to society.”¹¹⁶ The report identifies four core principles of “Explainable AI”: (1) “Explanation,” or whether the system provides the evidence that it relied upon for its outputs; (2) “Meaningful,” or whether that evidence or data is understandable to the user or recipient; (3) “Explanation Accuracy,” or whether the explanation actually reflects how the output was generated and the metrics used to make that judgment; and (4) “Knowledge Limits,” or whether the system produces outputs within the parameters of its design and training and whether those outputs reach the intended or appropriate confidence level.¹¹⁷ The study concludes by noting that there are some aspects of decision making and explanation that humans are better at than AI-empowered machines and some that algorithms are better at than humans.¹¹⁸

As with so many aspects of AI, the questions for policy makers and legal decision makers, like judges, are firstly, when to use AI at all for a particular task, and secondly, when using AI, how to structure human-machine collaboration so as to maximize the benefits of AI and minimize and mitigate its risks, including the risks of bias and lack of transparency. We call this latter challenge the Centaur’s Dilemma, after the mythical animal that is half human and

114. P. Jonathon Phillips et al., NISTIR 8312, Four Principles of Explainable Artificial Intelligence 12 (Sept. 2021), <https://doi.org/10.6028/NIST.IR.8312>.

115. *Id.* at 13.

116. *Id.* at 1.

117. *Id.* at 2.

118. *Id.* at 18–21.

half horse as well as the Department of Defense adoption of the “Centaur Model” to describe human-machine teaming. Here the centaur is part machine and part human and the dilemma is to determine how much human choice and control to require and how much autonomous machine choice and control to require (or permit) of each AI application, output, or decision. Three related questions for judges and experts to ask in the context of AI-generated evidence are:

- whether a transparent and explainable methodology was used, and if not, why not;
- whether the methodology used has been demonstrated to be effective in the context for which it is offered; and
- whether the AI-generated evidence is understandable to the factfinder for the purpose for which the AI output is being used or offered.

Judicial Roles and High-Level AI Takeaways for Judges

Judges will play at least five roles related to AI in the courtroom. First, judges will oversee the process of discovery in civil and criminal cases: where AI outputs are at issue or may be offered as evidence, judges will need to determine how much discovery to permit into the AI’s design, background data, and accuracy. Second, they will serve as evidentiary gatekeepers, applying the Federal Rules of Evidence (or state equivalents) to proffers of testimonial and documentary evidence, including and perhaps especially Rules 401, 402, and 403. Third, judges will serve as guardians of the law (some may prefer the phrase “interpreters of the law”), protecting and preserving the rights and values embedded in the Constitution as well as statutes and rules of procedure and evidence. Fourth, judges may serve as potential AI consumers who need to decide whether to receive, review, or rely on AI-generated outputs to inform bail, probation, or sentencing decisions. Some risk-assessment tools rely on predictive algorithms, raising legal concerns about due process, equal protection, and confrontation, triggered in part by technical concerns about accuracy and bias. Fifth, judges will serve as communicators, translating the sometimes complex inputs behind AI into plain-language instructions for jurors and case law precedent for lawyers.

The previous sections introduced the technology behind AI. This section highlights nine technical aspects about AI of which judges should be aware in their roles as referees, gatekeepers, guardians, potential consumers, and communicators. The following sections address bias, predictive algorithms, algorithms used by lawyers, discovery, and AI as evidence.

There Are Many Different Methodologies and Designs

Because there are different AI methodologies, each application should require authentication and validation, not just in concept but as applied in each context and potentially moment in time. As described above, within the category of machine learning, there are multiple ways to teach a machine to learn using data,¹¹⁹ the most common being supervised learning, unsupervised learning, and reinforcement learning. Any of those might make use of artificial neural networks employing deep learning with several or many hidden layers.

Yet there are still other theories and methods for teaching computing machines to learn, each built into the operative algorithm. These alternatives include evolutionary or genetic algorithms, inductive reasoning, computational game theory, Bayesian statistics, fuzzy logic, hand-coded expert knowledge, and analogical reasoning.¹²⁰

In addition to deciding on the learning methodology, computer engineers must also decide how much depth and breadth to apply to any deep learning neural network—in other words, how widely the algorithm will search (breadth, also referred to as width) and how many layers of internal inputs and outputs it will employ before providing an output (depth). With facial recognition, for example, breadth might represent the number of datasets an algorithm searches. Depth might be illustrated by the number of points on a human face the algorithm is programmed to analyze before providing an output. Increases in network size tend to be required to capture the complexity of modern AI algorithms. Such increases create a challenge, however: the greater the depth—the number of layers in the neural network—the harder it will likely become to determine which factors were determinative in the output prediction. This could become important to the extent there is risk or concern that bias or some other factor might undermine outcome accuracy.

Some algorithms are designed to provide outputs, plural—for example, a range of match faces with a facial recognition algorithm, or a range of products with a shopping recommendation algorithm. As mentioned above, given the black box problem of deep learning, there are additional methodologies engineers can incorporate or substitute to make an algorithm's internal calculations more transparent, or fully transparent, to users.

A court will need to satisfy itself that the specific AI application (as opposed to AI generally), its design, and its specific use meet the foundational requirements for the purpose for which it is being offered into evidence or used by a court. Verification will entail considering the theory and methods behind the AI,

119. *Id.* at 3.

120. See The Weaponization of Increasingly Autonomous Technologies: Artificial Intelligence 5, United Nations Institute for Disarmament Research (UNIDIR) No. 8 (2018), <https://perma.cc/HQM5-BHWQ>.

including the nature of the datasets used to train, test, and validate the AI, as well as, in the case of deep machine learning, inquiring into the design of the neural network.

Most AI Is Iterative and Should Be Tested and Validated Continuously

Many AI/ML models learn, for better or worse, as they proceed.¹²¹ That means AI systems need to be tested and validated on an ongoing basis. In other words, if a machine is learning, its variance rates and accuracy should change as well—in theory, in the direction of greater accuracy and lower variance, but, as described above, the machine’s accuracy and reliability may degrade.

If the underlying data are biased or poorly selected in that they do not match real-world conditions, that may undermine the accuracy of the AI or embed bias in the AI’s application. Results from the testing phase may not match actual outputs in the field. In context, judges, experts, and litigators will have ample reasons to test the reliability of any AI evidence offered in court, and judges will need to determine in context just how wide to open the door to expert testimony and discovery about matters like datasets, algorithmic design, search parameters, bias, and neural network architecture.

Humans Are Always Involved

Machines do what they are programmed to do, not because they choose to do so, but because they are programmed to do so—including learning on their own. Software drives machines. And humans, in the first instance, write software and design programs. Behind each AI application there are human choices, human values, and human bias that may impact the operation of the algorithm and the accuracy of its results. Humans select not only the data but also the metrics the algorithm uses to frame and analyze that data.¹²²

In the operation of AI, humans are also involved. Under current vernacular in the AI field, humans are said to be “in-the-loop,” “on-the-loop,” or “out-of-the-loop.” As implied, in-the-loop describes humans in functional control of an application, deciding when and how it is used. On-the-loop describes humans observing AI but not controlling it, but with the option to do so. Out-of-the-loop describes an autonomous or semiautonomous system operating

121. Other models might learn during training but are not capable of further learning on their own during deployment. Still others might be trained and later adjusted or retrained only with human help or supervision.

122. Remarks of Nisheeth Vishnoi, *supra* note 22.

automatically. These terms are imprecise in at least two regards. First, they describe a wide variance of conduct within each category and thus may convey a sense of control and oversight that is, in operation, absent. More to the point, they are insufficiently descriptive to apportion accountability and responsibility for the purpose of legal judgments. Take the example of a “driverless car.” Some driverless cars are configured to employ a safety driver as an observer or, in the case of a semi-driverless car, a driver with shared responsibility for the operation of the vehicle. Other driverless cars, without a human in the car, may operate under remote human control. In each of these three scenarios, at any moment in time the vehicle may be driving autonomously without human control, it may be following the explicit direction of the remote or present driver, or the human driver may be keenly observing the operation of the vehicle without overriding the car’s computers. In each case, humans were out of, in, and on the loop.

However described, a human is always involved with an AI application. For courts, the factual questions will be: Who designed the seed algorithm? Using what metrics or weights? Who trained the algorithm? Using what data? Who collected the data? Who validated the data? Who used the algorithm or monitored its use? These factual questions will lead to legal questions. For example, where *Crawford*¹²³ applies, multiple persons might be called as witnesses regarding the design and operation of an AI algorithm. Because humans are always involved with AI, there will be persons who can, if relevant and material, provide answers to the sorts of questions essential to authenticating and validating the use of AI:

- What is the AI trained to identify, how has it been weighted, and how is it currently weighted?
- Does the system have a method to transparently identify these answers? If not, why not?
- Are the false positive and false negative rates known, if applicable, or hallucination rates? If so, how do these rates relate to the case at hand?
- How has AI accuracy been validated, and is the accuracy of the AI updated on a constant basis?
- What are the AI’s biases? (See “Probing for Bias” questions below.)
- Is authenticity an issue?
- How do each of these questions and answers align with how the AI application is being used by the court or proffered as evidence?

Judges might also consider that a qualified AI expert or witness ought to be able to credibly answer these questions, or perhaps the expert or witness may not be qualified to address the application at issue.

123. *Crawford v. Washington*, 541 U.S. 36 (2004).

AI Predicts, It Does Not Conclude

AI is generally a predictive tool based on statistics. Through weighted calculation an algorithm predicts an outcome. A facial recognition algorithm, for example, might predict the likelihood that one photograph in a set of fifty returned by the machine as potential matches is indeed the correct match. What the algorithm does not do is confirm that an image presented is in fact a match in the same way that a chemical test confirms the presence of a compound. This is one reason why engineers use the term “confidence threshold” in describing the accuracy of an application. The FBI facial recognition algorithm, for example, is designed not to conclusively find a match but to find pictures that might match. As the 2019 GAO report on the subject stated, the algorithm is most accurate when offering a range of potential matches.¹²⁴

In the case of a Google search algorithm, for example, the algorithm is predicting that one of the provided links will respond to the query. This is self-evident if one asks a question like, “Who was George Washington?” The algorithm is likely to provide a Wikipedia link to a webpage about the first U.S. president. It is also likely that many readers will conclude that the algorithm has *answered* the question: “George Washington was the first president of the United States.” It has not. The algorithm has predicted that one or more of the links provided will answer the question, and likely in descending order of probability as the links are listed. Modify the question a bit, and the predictive aspect becomes more evident. If you ask, “Who is my friend George Washington?”—a quite different person than the first president—Google responds with sites listing the first president’s friends. That is the algorithm’s best prediction as to which links will answer the question based on code matching, likely use of the word “friend,” and the way prior readers have responded to similar word searches. In other words, like a shopping algorithm, the search algorithm is tracking whether the searcher “bought” the response by clicking on it and measuring how long the searcher stayed. Of course, it has not answered the question at all and is nowhere near to providing a link that will answer the question about the user’s friend George Washington—not without more details that can help shape the predictive outcome.

In a medical context (perhaps coming before a court in a malpractice case) an input might query, “Is this a picture of a benign or malignant tumor?” To respond to that question, an algorithm trained on prior pictures of tumors might break the picture into quadrants and subcomponents, as a facial algorithm might do, and then compare the picture submitted to database images of tumors. Based on all accessed images of benign and malignant tumors, the algorithm will predict whether the picture is a better match for one or the other. What the algorithm offers that a human does not is the capacity to search multiple databases rapidly for comparative patterns, as well as the ability to break the image into

124. See Harwell, *supra* note 44.

subordinate components in a way humans cannot, and thus to see connections and patterns the human eye cannot. Moreover, the algorithm is neither affected by fatigue nor subject to ordinary human distractions, pressures, and emotions. If the algorithm has not been trained properly, or trained to identify new patterns, it is less likely than a human to identify a rare disease or new manifestation of an existing disease, raising the prospect of a false negative. One can imagine in a malpractice case how the parties might litigate the manner in which any human-machine teaming occurred. Where a tumor was not diagnosed, a plaintiff might argue that doctors placed unreasonable reliance on a “negative” AI output. Alternatively, a plaintiff might argue an unreasonable lack of reliance if a broader use of AI databases was not employed.

For sure, algorithms are used to “make decisions” as well as to predict outcomes. Algorithm-generated credit scores, for example, are used to determine eligibility for loans and credit cards. The algorithm produces the score, but it is humans who decide what parameters should inform that score and it is humans who decide what score should be used as a threshold cutoff for loan eligibility or rates. The algorithm in essence is making a numeric prediction about the reliability of the loan; the human is making a decision to rely on the prediction and at what confidence level to do so. Likewise, driverless cars make “decisions” all the time. However, they do so based on human programming and values embedded in that programming, a centaur model of decision, even in a fully autonomous vehicle.

Currently, Accuracy Often Depends on the Volume and Quality of Data

If an algorithm has only been trained on one picture of a cancerous tumor or has never seen a cancerous tumor, then it will be less likely to correctly identify a tumor in response to a query. This is an important limitation on the capacity and accuracy of current AI. Moreover, volume here is not measured in hundreds, but in hundreds of thousands of pictures. A human performing the same task with only one picture will more likely identify a tumor using intuition, judgment, and experience, as well as external factors the algorithm cannot assess, like the patient’s unique pain threshold or situational responses to touch and feel. As discussed above, today’s large language models (LLMs) train on vast datasets, though researchers are developing other types of modeling less dependent on large datasets, such as information lattice models.

The quality of data is also important and will likely remain so. Data can be flawed in a variety of ways. Dated data, known as stale data, are more likely to generate inaccurate results (“temporal bias”), as are incomplete or underrepresentative data (“population bias”). A facial recognition algorithm trained on driver’s license pictures or parole pictures is more likely to identify pictures reflecting the demographics represented in the databases. This has the potential

to increase the false negative rate for underrepresented groups and to increase the false positive rate for overrepresented groups.

Likewise, data may possess flaws that impact algorithms but not humans. Algorithms may discern links in data or perceive patterns in data creating matches, based on elements or numeric formulas that are unintended or that humans would not discern. In our bank robber scenario, the algorithm may match digital sequences found in the image based on irrelevant factors, such as a common backdrop in a photo or pattern on the robber's face mask. In either instance, there may be a match but not a meaningful match. If this occurs within the neural network, it may skew a result in a manner unseen and unknown to the user. The output is a face, but the user may not know this face has been passed through to the output stage because of similarities in the picture backdrops, not the face itself, because AI systems often are unable to explain their results.

Some AI Is Not Explainable but More Transparent Methodologies Are Being Developed

The black box of neural networks renders many AI outputs unexplainable. However, the field of explainable AI is expanding as is the development of tools allowing reconstruction of the reasoning behind AI outputs. This capacity has many legal implications, such as the ability to reconstruct or explain AI results that are biased, incorrect, or result in accidents (driverless cars) or mistakes (medical diagnoses). Therefore, users of AI, including proponents of its admission into evidence, should articulate what is explainable and not explainable and why, based on the tools available to do so at the time of use or at evidentiary proffer.

As will be seen below, this limitation makes certain predictive algorithms particularly susceptible to error. It is essential that judges and factfinders understand the ways data and design can embed witting and unwitting bias, as discussed in the section below titled "AI Is Biased," impacting the predictive accuracy of AI.

The Heart of AI Is the Algorithm

If the accuracy of an AI application often depends on the amount of data on which it is trained, it depends even more on the algorithm applied to that data. As noted previously, an algorithm is a process that guides the software determining which data are selected and how they are weighted. The *choice* of algorithm is a choice of decisional metrics or framework—an analytical lens or value-laden perspective.

Using the analogy of an algorithm as the recipe a chef uses: the chef chooses not only the end-dish but also dietary restrictions (vegetarian, low sodium),

flavor profile (sweet, acidic, spicy), cultural heritage, ingredient measurement system (metric, English), etc. As this is AI, not a simple algorithm, the chef supplements her recipe every time she cooks in a way that only she knows. What is more, the sous chefs supplement the recipe when no one is looking. Thus, in some cases no one can be quite sure what gives the recipe its distinctive taste; and, if the chef knew, she would not tell because she wants customers to continue to come to her restaurant. Restated, if one knew the Google search algorithm, every search platform could, in theory, be as good, provided of course, that the algorithm could access and apply the same level of data (Google's data) with which to train, test, validate, and apply the algorithm.

Thus, the heart of many disputes about the use of AI-generated evidence in court or the use of AI tools to inform judicial decision making will revolve around access to and disputes over the accuracy and quality of algorithms. This may be the proprietary secret AI companies want most to protect, because it is the recipe to their market success and because too much inquiry may undermine confidence in the AI's capacity.

Here are several questions judges should contemplate before using an AI application or admitting one into evidence:

- To what extent will the court allow parties to discover the content of an algorithm? Will discovery include the data on which the algorithm was trained, tested, and validated?
- In the context presented, do due process, equal protection, or the right to confrontation require access to an underlying algorithm or its supporting training, validation, and use data, or to the human designers, programmers, and users?
- If discovery is permitted, what safeguards, if any, will the court use to protect the proprietary value of the discovered information? In the context of criminal risk assessments, do rights to due process, equal protection, or confrontation require the algorithm and underlying data to be made public?
- To what extent is discovery necessary to apply *Daubert*? (See below.)
- In the context presented, can ex parte and in camera judicial review adequately and legally substitute for public adjudication? Or should the parties or the public have access to the algorithms and data?

These questions might lead to further questions:

- Will the court or a jury be able to understand the underlying technology, and is such understanding necessary for a fair adjudication of the facts?
- If so, what is the appropriate mechanism to provide that understanding?

Narrow AI Is Brittle

As noted earlier, narrow AI is not particularly good yet at situational awareness. The driverless car may not timely identify novel objects on the road. Judges will therefore have to consider whether the scenario or fact for which an AI application is offered presents questions involving situational awareness. If so, they should then ask in what manner the algorithm, the data, and the training are keyed for such circumstances and whether the accuracy rate varies in such contexts. They might ask:

- Is the AI in question brittle?
- What is the error rate, i.e., the rate at which the AI reaches the wrong (false positive or false negative) result?
- Does the AI perform in the same manner and with the same accuracy in “real-world” operation as it does (or did) in lab conditions and testing?

AI Is Also Nimble

Proponents of AI tend to emphasize its strengths, opponents its weaknesses. Of course, the strengths and weaknesses of any AI application must be assessed on a case- and application-specific basis. One strength of AI, however, is its general capacity to identify, aggregate, and derive meaning from data in ways that humans cannot. With driverless vehicles, for example, this capacity is simply illustrated by the fact that, with the right sensors, driverless vehicles can “see” in all directions at once and calculate, with mathematical precision at speed, the distance needed to brake. AI can also identify patterns, anomalies, and links in data that humans cannot. In many cases, AI is better than humans at tasks like comparing pictures of tumors to database images of benign and malignant tumors. And AI can do all this instantaneously and, in many cases, more consistently than humans, a point illustrated with reference to comparisons between humans reviewing documents for security classification or discovery purposes and algorithms doing so. Different humans often reach different conclusions. A good AI will be more consistent, albeit it might be more consistently correct or incorrect in its result. Of course, depending on the context, humans will need to determine whether the patterns and links that are made are relevant and reliable for the purpose presented.

Generative AI Can Hallucinate

AI can create false content, as described above.

AI Is Biased

AI, like humans, has biases. It is also an area of AI that will attract adjudication, as the next section will suggest.

Bias

Bias is often associated with the human application of stereotypes or prejudices to an ethnic, gender, racial, or other identity group. In U.S. law, such categories are generally recognized as “suspect classes” in equal protection law under the Fifth and Fourteenth Amendments.

As judges know, any application of law that treats classes of persons differently from the populace as a whole, if challenged in court, must pass either a strict scrutiny test, intermediate test, or rational basis test, depending on the suspect class. An application of law that is facially neutral but adversely affects one protected group more than another might also be subject to a disparate impact claim. For example, a hiring algorithm that disproportionately favored one group over another might be subject to a disparate impact lawsuit.

However, when technologists consider bias, they are often referring to the variance between what an AI is intended to predict or perform and the accuracy with which the algorithm predicts or performs that function. This “algorithmic bias” might include social biases incorporated into the AI’s design and implementation—but with AI, bias is usually defined more broadly as a witting or unwitting (conscious or unconscious) predisposition that can undermine the accuracy of an AI application or output. Bias thus addresses a range of cognitive tendencies that can adversely affect objective analysis and technical accuracy. Significantly, AI bias also incorporates and describes unintentional design and data flaws that can impair the accuracy of AI outputs. Because humans design AI algorithms and choose the data that trains the software, developers’ choices can incorporate varying types and degrees of bias into the algorithm’s design. Unintentional bias is often difficult to discern because it is embedded in the design of an AI system or in the data used to train an algorithm. Decision makers may subsequently place undue reliance on AI outputs because they do not know to what extent, if any, the outputs are predicated on biased input, e.g., a facial recognition algorithm or a predictive algorithm trained on limited or poor-quality data.

When it comes to algorithmic bias, there are four immediate takeaways for judges:

1. Judges (and the law) use the term *bias* in a different and more specific way than computer engineers. For AI specialists, algorithmic bias refers broadly to the difference between an algorithm’s output and the desired outcome, not necessarily to bias of the sort addressed by the equal

protection clause. Algorithmic bias can be caused by human prejudice of the sort courts typically address, cognitive bias of the sort behavioral scientists typically address, design and data flaws of the sort computer engineers address, or all of these (admittedly overlapping) categories.

2. *There is no such thing as a bias-free algorithm.* There is a tendency to believe that “numbers are neutral” and present objective truths, but numbers may produce erroneous or inaccurate results.¹²⁵
3. Through careful engineering, thoughtful use of data, and adjusted algorithmic weights, it is possible to create AI systems with lower margins of error.¹²⁶ It is also possible that reducing one form of bias by adjusting, for example, the underlying analytical framework or datasets can allow other forms of bias to creep in.
4. Judges, as evidentiary gatekeepers, can mitigate or block the use of weak or biased AI by asking the right foundational questions. Knowing what to ask starts with an understanding of the forms that algorithmic bias might take.

Forms of Algorithmic Bias

The United Nations Institute for Disarmament Research suggests several categories and sources of algorithmic bias.¹²⁷ Starting with the Institute’s findings, we highlight eight forms of potential AI bias: statistical bias, moral bias, training data bias, inappropriate focus, inappropriate deployment, interpretation bias, unwitting human bias, and intentional bias.¹²⁸ We also discuss the issues of overfitting and outliers as potential sources of bias.

Statistical bias

Statistical bias might occur when an algorithm’s predicted outcomes deviate from a statistical standard, such as the actual frequency of real-world outcomes.¹²⁹ This deviation can be caused by bad statistical modeling or incorrect or

125. Joni R. Jackson, *Algorithmic Bias*, 15 J. of Leadership, Accountability & Ethics 55–65 (2018), <https://doi.org/10.33423/jlae.v15i4.170>.

126. Jake Silberg & James Manyika, Notes from the AI Frontier: Tackling Bias in AI, McKinsey Global Institute (June 6, 2019), <https://perma.cc/K8Z8-WUAE>.

127. *Algorithmic Bias and the Weaponization of Increasingly Autonomous Technologies: A Primer*, UNIDIR Resources No. 9, United Nations Institute for Disarmament Research (UNIDIR) (2018), <https://perma.cc/N2H7-DWPT>.

128. This section relies heavily on our prior work with co-author Matthew Mittelsteadt, *An Introduction to Artificial Intelligence for Federal Judges*, *supra* note 8.

129. *Id.* at 2.

insufficient data. The difficulty of calculating the infection or mortality rate of a disease such as Covid-19 illustrates the problems that can result. At the outset of the pandemic, AI-driven modeling of infection rates differed widely, in part because the models could not account for those who had the disease but did not show symptoms. Thus, it was only within closed data samples, such as the passengers aboard a cruise ship, that the models could account for an asymptomatic pool (because all the passengers were tested before they were allowed to leave the vessels). The risk with cruise ship findings was having too small a sample pool, and one not necessarily random or representative of a cross-section of the population in more typical, land-based living conditions.

Moral bias

Moral bias occurs when an algorithm's output differs from accepted norms (regulatory, legal, ethical, social, etc.).¹³⁰ For example, an algorithm may weigh factors that the law or society deem inappropriate or do so with a weight that is inappropriate in the context presented. A hypothetical predictive crime algorithm might use data derived from the current prison population to “predict” rates of recidivism. Given the disproportionate number of people of color imprisoned, the algorithm might produce biased results by explicitly or incidentally weighing “race” and ethnicity or their proxies, and perhaps do so within the black box of a neural network. AI neither thinks nor understands the world like humans, and unless instructed otherwise, its results can reflect an ignorance of norms found in the equal protection and due process clauses.

Training Data Bias

Like humans, AI learns from experience; however, AI experience is based exclusively on data, often hand-selected by a human developer. Inaccuracies or misrepresentations in this data can perpetuate biases by embedding them in algorithmic code or weights.¹³¹ In other words, the results are skewed; the algorithm produces wrong answers. As presented in the discussion of the machine learning life cycle, above, biased data might be outdated (“temporal bias”), under- or over-representative of certain groups (“population bias”), or simply reflective of individual or systemic social prejudices. For example, an algorithm intended to identify potentially successful job applicants might rely on past successful job performance as an indicator of future successful job performance and derive from that data certain preferred hiring characteristics like age, school, and experience.

130. *Id.*

131. *Id.* at 2–3.

But if the data are from a period when women or other marginalized groups were not numerically well represented in the relevant employment market or educational pool, at least 50% of the potential workforce might be excluded from results. Such data might likewise incorporate human bias in the form of a past company policy to only hire persons from certain schools. The criterion might have seemed objective when the company adopted the policy, but it necessarily incorporates the socio-economic and other biases of the college admissions processes of the time. Thus, the algorithm might exclude candidates as good or better than those from whom the dated dataset was sourced.

Similar concerns have been raised about algorithms designed to inform parole decisions by predicting recidivism risk. These algorithms, critics argue, appear¹³² to rely on socioeconomic status, neighborhood location, and past crime statistics as predictive criteria of future criminal conduct, potentially resulting in a self-fulfilling prediction with disparate racial and socio-economic effect. Criminal risk assessments used in bail and sometimes sentencing meet with analogous criticisms, discussed below under Algorithms That Predict Human Behavior. Proxies for suspect categories, such as zip codes or last names, might also introduce bias into data, as discussed above under How Machines Are Trained to Learn: The Machine Learning Life Cycle. Proxies may be hard to detect. Additionally, data used to test or validate algorithms might also infuse bias.

Inappropriate Focus

This occurs when an algorithm's training data are ill-suited to the algorithm's task.¹³³ This might lead the algorithm to identify factors within a neural network that, though objectively reasonable, are logically irrelevant to the desired outcome. An algorithm that matches faces based on colors, backdrops, or lighting demonstrates inappropriate focus bias.

One can readily imagine how similar bias might migrate into datasets designed to train machine-learning AI to predict terrorism recruitment or threats. To begin with, the data may rely too heavily on international versus domestic actors owing to the designer's perceptions or the selection of training data. And because the amount of data may be limited (in contrast to, say, an Amazon or YouTube algorithm), human actors may put too much credence in the reliability of the predictive output. In general, the more data used to train a predictive algorithm, the more accurate the result. An algorithm trained to predict terrorism risk based on a stereotyped "terrorist profile" will, unsurprisingly, be best at locating persons who meet that profile. More persons with the profile

132. They are often proprietary, as was the case for the risk assessment tool at issue in *Loomis v. Wisconsin*, 881 N.W.2d 749, 759 (Wis. 2016), *cert. denied*, 137 S. Ct. 2290 (2017).

133. Baker, Hobart, & Mittelsteadt, *supra* note 8, at 4.

will be identified as potential terrorists, and more may be found to be engaged in suspicious activities because of increased scrutiny, thus appearing to validate the algorithm and the choice of criteria. Potential terrorists who don't fit the profile may be erroneously omitted.

As the example demonstrates, the risk is not just in false positives, which is the focus of much bias analysis to date, but in the potential failure to identify credible risks: false negatives. Disparities in facial recognition data between males and females or people of different ethnicities could lead to greater inaccuracies in identifying female subjects or people of color, increasing the number of false positives—for example, the number of innocent people selected for extra screening or questioning at airports. In contrast, an inability to identify known subjects or threats—for example, a missing or wanted person or Amber Alert kidnap victim on CTV camera feeds—has security implications.

Inappropriate Deployment

This happens when a system is used in a context for which it was not designed, tested, and validated.¹³⁴ For instance, a driverless car trained for driving in the United States might not be able to handle left-hand driving in the United Kingdom. A human (we hope) would adapt to such a change; a driverless car algorithm would need more training.

Interpretation Bias

This occurs where an algorithm's output is confusing or subject to incorrect interpretation by those working with the technology.¹³⁵ Users of facial recognition technology might expect singular, or perfect, matches, in contrast to what most facial recognition algorithms—including the FBI's—actually produce, which is an array of potential matches, none of which might represent the individual sought, leaving the interpretation and conclusions to human users.

Interpretation bias can also occur because of ambiguity embedded in the algorithmic design—for instance by software designers who, unaware of cultural or linguistic cues, overlook or misuse phrases and concepts, skewing results. Sometimes the reasoning behind a match is necessary to understand its value or import. Engineers might design algorithms to search for particular words or phrases, with the goal, for example, of identifying persons engaged in radicalizing internet users. Insufficient knowledge of culture and language, however, could have unintended consequences. Phrases like “the bomb,” “knock 'em

134. *Id.*

135. UNIDIR Resources No. 9, *supra* note 127, at 5.

dead,” and “kill it,” all can mean something in American vernacular quite different from what might be intended in a terrorist cell. By the same token, an algorithm designed by an engineer who, for example, does not know the import of “the fourteen words” (which form two slogans of white supremacists) may inadvertently enable a potential data threat stream to escape detection.

Unwitting Human Bias

This refers to the unintentional infusion into an application of human preferences, stereotypes, values, fears, or knowledge. Consider an algorithm intended to predict risk. An engineer might apply engineering principles to a risk equation. But what is risk? An algorithm will almost certainly incorporate the particular fears, risk tolerances, and perceptions of its designers. (The problem may be compounded when the algorithm is both human and machine generated—a “centaur”—clouding where and how bias might have entered the system.) But the algorithmic equation does not account for human behavior, which is informed not only by the calculation of objective zero-sum costs but also by the emotional impact of fear.

Use of racial, gender, and other social descriptors in algorithms is inherently risky and potentially fraught with ethical and legal issues. One can imagine how both intentional and unintentional human bias might enter the equation as a computer scientist embeds what he or she believes are parameters associated with a “race” or ethnicity into facial recognition software. Racial and ethnic categories are inherently ambiguous social constructions covering wide continuums of individuals. Similarly, one can see how nuance might “fool” an algorithm intended to identify age based on the subtle distinctions of faces alone without allowing for the possibility of make-up or efforts at disguise. Bias may also occur unwittingly in machine-learning applications that may not be designed to depend on social identity descriptors but nonetheless rely on such characteristics within the neural network black box. Bias can lead to both the under- and over-inclusion of the targeted group, as “race,” ethnicity, gender and other social categories are malleable concepts. Scholars like Charles Szymański of the University of Białystok note the risk of unintended age bias, as occurred with the use of an employment screening algorithm that sorted candidates based on whether the candidate submitting an application utilized the default browser or had downloaded a substitute browser. The tool, which was intended to identify technology-savvy applicants, had the unintended effect of excluding candidates based on age.¹³⁶ In effect, the choice of browser served as a proxy for age.

136. Charles F. Szymański, *The Future of Labor Law*, Presentation at Trnava University, Slovakia (Oct. 5, 2022).

Intentional Bias

Scientists, operators, and decision makers may use AI facial recognition tools or predictive algorithms to target disfavored or vulnerable groups. Algorithms can be designed to identify and select certain real and perceived social descriptors associated with “race,” gender, sexuality, national origin, religion, disability, and more. Facial recognition technology can identify and track certain ethnic groups as is the case in China with “Uyghur characteristics.” Clearly pernicious in the profiling of Uyghurs (or more accurately, a band of physical characteristics Chinese state security services associate with Uyghurs), such technology presents one legal question that may arise for judges: Whether, when, and under what controls is the purposeful use of social identity descriptors ever an appropriate search parameter?

On the one hand, there are qualitative differences between the reactive versus predictive uses of social identity descriptors. For example, using an individual suspect or victim description, including descriptors like “race,” gender, and age, in response to a credible, individualized predicate, in context, might make sense. Of course, one needs to consider that the initial social indicator that might trigger the use of an AI application may itself be affected by cognitive, societal, or the personal bias of witnesses. In the context of using facial recognition to search for a known suspect or a person identified in an Amber Alert, one would not necessarily expect law enforcement to employ “race”-, gender-, or age-neutral input or an algorithm incapable of searching for the specific or reported characteristics of the suspect or victim. Using individualized, known suspect descriptions as an “identifying factor” is “perhaps the least controversial use of race in the context of the Fourth Amendment” under federal case law.¹³⁷ However, as already noted, one potential challenge to individually based suspect descriptions is that to the extent social identities are fluid rather than fixed, they may be difficult to code, and they may be difficult for officials to implement fairly and accurately in AI and in general contexts.

On the other hand, using a suspect category and accompanying social identifiers to identify and predict persons who might engage or have engaged in an unlawful act—rather than relying on individualized, behaviorally based reasonable suspicion or probable cause—is an exercise in bias. Courts that have not “sidestepped” the issue have generally rejected explicit racial profiling based on “incongruity” arguments (that a particular person is “out of place” in a particular location or neighborhood).¹³⁸ Likewise, with exceptions in the immigration and border contexts, courts have rejected “propensity” arguments (that membership in a particular social group suggests a “propensity” for an individual to commit or have committed a crime).¹³⁹

137. *Farag v. United States*, 587 F. Supp. 2d 436, 462 (E.D.N.Y. 2008).

138. *Id.* at 462–63.

139. *Id.* at 463–64.

Courts may need to consider whether an AI application might blur the lines between individual suspect identification and group profiling. An AI application—for example a facial recognition database—might cast a wider net than a traditional human law enforcement investigation, to the point where what started as an individual suspect identification begins to look more like group profiling, opening more people (all of whom or all but one of whom might be innocent) to suspicion and investigation.

Overfitting And Outliers

As Matt Mittelsteadt has identified,¹⁴⁰ two significant risks posed by the use of machine learning in criminal risk assessments and other legal contexts are “overfitting” and “outliers.” Overfitting occurs when the ML model is too tailored to the data it has been trained on and does not account for ambiguities or variations.¹⁴¹ Generally in machine learning development, this problem is addressed by ensuring that the data the ML algorithm is trained on are separate from the data it will encounter in use (provided that the training data is still representative of the data that will be encountered in use). A model thus “generalized” should be flexible enough to correctly interpret data it has not encountered.¹⁴² If this data separation is not made, biased results may occur.

As discussed below, some courts have already used algorithms in sentencing. Overfitting, like other biases and weaknesses of AI, presents particular risk in sentencing and other applications where liberty interests are at stake. If a hypothetical ML sentencing algorithm is built on a training set of past offenders, the AI could design its neural network with results custom fit for those specific offenders. If a person reoffends, perhaps with a lesser crime, and his data are used to train the algorithm, there is risk that in calculating a sentence the algorithm might find and match his prior personal data and reproduce the prior sentence; statistically, the sentence will be the best match for his case. In essence, the algorithm might conclude, within its black box, “For someone with this background, we give a sentence of X.” In effect, the sentencing algorithm has shown a focused bias targeting a specific person (or even a person with similar personal data unrelated to any offense; this overlaps with population bias). (Note this is only a hypothetical model: we do not know of any risk assessments currently in use that are designed to correlate sentences with crimes for which an individual has already been convicted, as suggested in the hypothetical above. Rather, in actuality, courts appear to be using risk assessments designed to predict *future* crimes or dangerousness (recidivism), designed for bail or post-sentencing purposes, and importing those

140. Baker, Hobart, & Mittelsteadt, *supra* note 8.

141. *What Is Overfitting?*, IBM (Mar. 3, 2021), <https://perma.cc/BN3F-GNKKJ>.

142. *See id.*

risk assessments into the sentencing context, raising constitutional questions about a defendant's right to individualized and accurate sentencing.¹⁴³⁾

Risk in the other direction might produce an “outlier,” which occurs when the input is sufficiently distinct from the scenarios built into training sets to confuse the algorithm. The situation is analogous to sentencing for a crime that is not included in the Sentencing Guidelines and is not readily analogous to an existing offense. Outlier inputs could lead to unpredictable, potentially biased, and incorrect results. New crimes that do not fit the algorithmic model for assessing bail or recidivism risk, or for which there is an exceedingly small dataset, could produce a similar result.

For sentencing, the threshold for not using ML or throwing out algorithmic results could reasonably be low, as the alternative—human decision making—is already the standard. In any event, it would seem incumbent on the proponents of using such an algorithm to demonstrate its validity and explain its functioning, just as a judge should (and in some jurisdictions is required to) put reasoning for a sentence on the record, allowing appellate courts and the parties to understand what occurred and why.

Probing for Bias

With AI as with people, some bias is always present. But steps can and should be taken to minimize bias and mitigate its influence. Mitigating bias starts with the design of an AI system and the nature and quality of the algorithm and the data on which the algorithm is trained, tested, and validated. An additional mitigator is sound process—timely, contextual, and meaningful. For policy makers and engineers, “timely” means at each point where input can directly influence outcomes, i.e., at the conception, design, testing, deployment, and maintenance phases of AI development and use. Sound process includes asking whether AI should be used for a particular purpose in the first instance—that is, whether an AI tool should be developed or employed for a particular use. “Contextual” means specific to the tool and use in question and with actual knowledge of its purposes, capabilities, and weaknesses. “Meaningful” means independent, impartial, and accountable. Specifically, the person using or designing an application should validate its ethical design and use. If a particular community or group of people is likely to be affected by the use of the tool, designers and policy makers should consult with that community or group in deciding whether and how to develop, design, or use that tool.¹⁴⁴ In addition, to the extent feasible,

143. See *State v. Loomis: Wisconsin Supreme Court Requires Warning Before Use of Algorithmic Risk Assessments in Sentencing, Recent Case*, 130 Harv. L. Rev. 1530, 1532–33 (Mar. 2017).

144. Baker et al., *supra* note 96.

the system's parameters should be known or retrievable. The system should be subject to a process of ongoing review and adjustment. The rules regarding the permissible use, if any, of social identifying descriptors or proxies should also be enunciated, clear, and transparent, and they should be subject to constitutional and ethical review.

Asking the right questions is crucial, not just for legal reasons but because the questions invariably underpin judgments about the reliability of the AI at issue. With the caveat that humans may not be able to understand or predict the behaviors of some AI systems with certainty or even reliably (which might counsel against using them at all or in particular contexts), here are some questions to ask to probe for bias:

- Who designed the algorithm at issue?
- What process of review was the algorithm subjected to?
- Were stakeholders—groups likely to be affected by the AI application—consulted in its conception, design, development, operation, and maintenance?
- What is in the underlying training, validation, and testing data? How has the chosen data been cleaned, altered, or assessed for bias? How have the data points been evaluated for relevancy to the task at hand? Is the data temporally relevant or stale? Are certain groups improperly over- or under-represented? How might definitions of the data points used impact the algorithm analysis?
- Do the data or weighted factors include real or perceived racial, ethnic, gender, or other sensitive categories of social identity descriptors, or any *proxies* for those categories? If so, why, and do they pass ethical and constitutional review? Have engineers and lawyers reviewed the way these criteria are weighted in and by the algorithm as part of the design and on an ongoing basis? In accord with what process of validation and review? (If these questions are too difficult for proponents of the AI to answer, is using the AI in a particular context too risky for constitutional rights, human safety, or other important interests?)
- Is there a disparate or adverse impact on the confidence threshold or is it otherwise based on racial classifications, ethnicity, gender, sexuality, ability/disability, nationality, and so on? If so, are there reasons for such disparity that survive constitutional and ethical review?
- Did the developer and user of the AI perform and retain disparate impact assessments before using the AI, or in monitoring the AI post-deployment?
- Is the model the state of the art? How does it compare against any industry standard evaluation metrics or application-specific benchmarks?
- How might the terms or phrasings in the user-generated prompts bias the systems' outputs? Can these prompts be phrased in a more neutral way? Do any of the terms used have alternative meanings?

- Are the algorithm's selection criteria known? Iterative? Retrievable in a transparent form? If not, why not?
- Does the application rely on a neural network? If so, are the parameters and weights utilized within the neural network known or retrievable? Does the design allow for emerging methodologies that provide for such transparency? If a transparent methodology is possible, has it been used? (If not, why not?) If transparency is not possible with state of the art technology, and a less transparent methodology is employed, what is the risk that the system will rely on parameters that are unintended or unknown to the designers or operators? How high is the risk? Is the risk demonstrated? How is the risk mitigated?
- Is the input query or prompt asking for a judgment, a fact, or a prediction? Is the judgment, fact, or prediction subject to ambiguity in response?
- Are there situational factors or facts in play that could, or should, alter the algorithm's predictive accuracy?
- Is the application one in which nuance and cultural knowledge are essential to determine its accuracy or to properly query it?
- Are the search terms and equations objective or ambiguous? Can they be more precise and more objective? If not, why?
- What is the application's false positive rate? What is the false negative rate?¹⁴⁵
- What information corroborates or disputes the determination reached by the AI application? Is the application designed to allow for real-time assessment? If not, is operational necessity the reason, or is it simply a matter of design? Is there a process for such assessment that occurs after the fact?
- Is the AI being used for the purpose for which it was designed and trained?
- Is the AI being used to inform or corroborate a human decision? Are humans relying on the AI to decide or to inform and augment human decisions?

Algorithms That Predict Human Behavior

Most algorithms are based on statistical prediction. In this sense, all algorithms are predictive. There exists a class of algorithms, however, that also seek to make predictions about future behavior based on past data. This happens all the time. Shopping algorithms seek to predict through data about prior purchases (past behavior) the predisposition of individuals to make additional purchases (future

145. Some AI models will not produce false positives or false negatives, e.g., a generative AI model that creates fictional or new artistic content.

behavior). YouTube uses algorithms that seek to predict additional videos a viewer might watch to generate additional views. They are called “recommendation algorithms,” but what they do is push products to viewers using predictions about their future viewing behavior based on past viewing behavior.

Judges may be called upon to determine whether particular algorithms that predict behavior, such as those in the criminal justice system, are constitutional or unconstitutional. As part of that determination, judges may need to examine whether the algorithm is accurate, and by what measure, and whether a particular application presents legal issues, as when, for example, a predictive algorithm embeds certain types of bias. Likewise, issues might occur because litigants are unwilling or unable to determine the parameters or datasets that informed an algorithm’s prediction and thus cannot reliably evaluate its accuracy as applied to the circumstances at hand.

Predictive algorithms are used, or might be used, in a variety of judicial and collateral settings. Some of the most frequently cited applications are in the criminal justice system. Many police departments and courts across the country use algorithmic risk assessments.¹⁴⁶ Police use such tools to predict where crime might occur and by whom.¹⁴⁷ Policing algorithms are not intended to predict individual conduct, though they might include the characteristics of individual actors in an area, like registered sex offenders. Proponents of such algorithms argue that algorithmic tools better focus finite police resources on areas where crime is most likely to occur, based on “neutral” data rather than the hunches, perceptions, or potential biases of police officers. The argument against them is at least twofold. First, such algorithms can generate their own reinforcing and circular logic. The algorithm predicts criminal conduct, police patrols are increased, and additional arrests occur, validating the accuracy of the algorithm. Second, the underlying data are not, in fact, neutral. Such algorithms may have a disproportionate racial and socioeconomic impact where they generate increased patrols in poorer neighborhoods with historically higher recorded crime rates and larger concentrations of people of color. In this way, they may also reflect past police practices and prosecutorial decisions focusing on communities and people of color. They may also have intentional racial impact to the extent they use “race” or socioeconomic status as predictive factors.

Some courts already use actuarial and/or algorithmic risk assessments in pretrial release, probation, and sentencing decisions; parole boards also make use of them.¹⁴⁸ Some of these algorithmic risk assessments are capable of machine

146. Randy Rieland, *Artificial Intelligence Is Now Used to Predict Crime. But Is It Biased?*, Smithsonian Magazine (Mar. 5, 2018), <https://perma.cc/88JS-VTS8>; *AI in the Criminal Justice System*, Electronic Privacy Information Center, <https://perma.cc/5PE6-9WQG>.

147. Rieland, *supra* note 146.

148. *AI in the Criminal Justice System*, *supra* note 146; Brandon Garrett & John Monahan, *Assessing Risk: The Use of Risk Assessment in Sentencing*, 103 *Judicature* No. 2 (2019), <https://perma>

learning,¹⁴⁹ a capacity that will increase with time. Private companies may develop the risk assessment algorithms; Northpointe developed the COMPAS system used by several states.¹⁵⁰ These companies may not release the underlying code for the algorithms (or training, testing, and validation data) for defendants to test and challenge.¹⁵¹

Using risk assessment algorithms to make or inform liberty decisions creates potential Fifth and Fourteenth Amendment due process and equal protection issues. It may also create Sixth Amendment Confrontation Clause questions. To identify a few due process issues:

- How may defendants meaningfully challenge the logic of an algorithm if the source code is kept from them? Is it enough for them to have access only to the inputs and outputs the algorithm processes and generates but not the decisional framework or processes it uses?
- How can a defendant meaningfully challenge an algorithmic risk assessment without access to its training, testing, and real-world-use data?
- If the algorithm uses machine learning, and no one, not even the developer, understands its “analysis,” how can courts ensure due process of law?
- How and should courts test algorithms for accuracy, especially when they predict future (i.e., unrealized) human behavior?
- If a particular risk assessment tool designed for one purpose, such as to predict future behavior (e.g., flight risk or recidivism), is applying that algorithm (along with the data it was trained on and the inputs its considers) for other purposes, such as sentencing for past behavior, is that consistent with due process?
- Should algorithms that depend on group data to predict future behavior be applied against individuals?

As with policing algorithms, any racial and other biases contained in or produced by risk assessments present equal protection issues. The adoption of risk assessment tools has caused controversy in this context,¹⁵² and there is a rich

.cc/BTC5-QWLE (describing and analyzing how some Virginia judges use and some do not use risk assessments at sentencing); see also Pennsylvania Commission on Sentencing, Guidelines and Statutes: Risk Assessment, <https://perma.cc/VJ9S-H9BZ> (“Act 95 of 2010 required the Commission adopt a sentence risk assessment instrument.”).

149. Danielle Kehl, Priscilla Guo, & Samuel Kessler, *Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing*, Responsive Communities Initiative (July 2017), <https://perma.cc/82NU-2K9D>.

150. *AI in the Criminal Justice System*, *supra* note 146.

151. See *id.*

152. For example, a 2016 ProPublica study determined that COMPAS was almost twice as likely to falsely identify a black person as a repeat violent offender as it was to falsely identify a white person as a repeat offender. The company contested this finding. Julia Angwin et al., *Machine Bias*:

academic literature debating (generally challenging) the efficacy and fairness of these tools.¹⁵³ However, there is yet little case law. Lawmakers or police departments using these tools might seek to replace, improve, or inform judicial decisions with “evidence-based”¹⁵⁴ algorithmic recommendations, or to decrease the incarceration rate by releasing more people before trial and during probation.¹⁵⁵ Critics argue that risk assessment tools not only have racially biased results but, through the ML process, may exacerbate racial inequalities in the criminal justice system. One significant concern is that using historical data to train risk assessment tools may only replicate and repeat at scale “the mistakes of the past.”¹⁵⁶ Over one hundred civil rights groups issued a joint statement detailing their concerns with pretrial risk assessments.¹⁵⁷ While the ideal of “evidence-based” practice may be appealing, the risk that an assessment tool may cause disparate treatment under a mantle of “data-driven” legitimacy warrants careful consideration. Whether any type of unbiased machine neutrality or fairness to all is possible is a matter of debate.¹⁵⁸ Some scholars have concluded that technical flaws in criminal risk assessments, in particular, pose grave concerns and cannot be fixed.¹⁵⁹

Academic commentary highlights several other risks of risk assessments: (1) that the existence of AI risk assessment tools will place pressure on courts to use the tools, whether they are accurate or not; (2) the psychological bias toward relying on empirical evidence more heavily than nonempirical evidence (“anchoring

There’s Software Used Across the Country to Predict Future Criminals. And It’s Biased Against Blacks, ProPublica (May 23, 2016), <https://perma.cc/K4ZP-2MP2>. See also Sam Corbett-Davies et al., *A Computer Program Used for Bail and Sentencing Decisions Was Labeled Biased Against Blacks. It’s Actually Not That Clear.*, Wash. Post (Oct. 17, 2016), <https://perma.cc/8UMU-8EK6>.

153. For an introductory overview of that literature, see *A Letter to the Members of the Criminal Justice Reform Committee of Conference of the Massachusetts Legislature Regarding the Adoption of Actuarial Risk Assessment Tools in the Criminal Justice System* (Feb. 9, 2018), <https://perma.cc/W3W5-YQ7V>.

154. *Loomis v. Wisconsin*, 881 N.W.2d 749, 759 (Wis. 2016), cert. denied, 137 S. Ct. 2290 (2017).

155. Derek Thompson, *Should We Be Afraid of AI in the Criminal-Justice System?*, The Atlantic (June 20, 2019), <https://perma.cc/4DKC-PDDL>.

156. Karen Hao, *AI Is Sending People to Jail—and Getting It Wrong*, MIT Tech. Rev. (Jan. 21, 2019), <https://perma.cc/MG6X-HQF6>.

157. The Use of Pretrial “Risk Assessment” Instruments: A Shared Statement of Civil Rights Concerns, <https://perma.cc/MS7U-R6KC>.

158. See, e.g., Christopher Bavitz et al., *Assessing the Assessments: Lessons from Early State Experiences in the Procurement and Implementation of Risk Assessment Tools*, Berkman Klein Center for Internet & Society 20–21 (Nov. 2018), <https://perma.cc/55FS-DRFY>; Craig Smith, *Dealing with Bias in Artificial Intelligence: Three Women with Extensive Experience in A.I. Spoke on the Topic and How to Confront It*, N.Y. Times (Nov. 19, 2019, updated Jan. 2, 2020), <https://www.nytimes.com/2019/11/19/technology/artificial-intelligence-bias.html>; Sandra Mason, *Bias In, Bias Out*, 128 Yale L. J. 2218, 2233 (2019).

159. Martha Minow, Jonathan Zittrain, & John Bowers, *Technical Flaws of Pretrial Risk Assessments Raise Grave Concerns*, Berkman Klein Center for Internet & Society (July 17, 2019), <https://perma.cc/PQY3-JK8A>.

bias”); and (3) that “most judges are unlikely to understand algorithmic risk assessments,” and therefore may misuse them or give them inappropriate weight.¹⁶⁰

Several generalized arguments for and against the use of predictive algorithms for human behavior emerge. Proponents might argue that predictive algorithms can identify patterns and trends humans cannot see and can do so rapidly, if not instantaneously, thus curtailing additional risk or harm. They might argue that predictive AI rests on the premise, and some would say reality, that neither judges nor law enforcement personnel can reasonably predict conduct based on judgment and intuition alone. AI simply has more data and excels at statistics. In the courtroom, predictive AI could add data to judgments about risk assessment, informing decisions on bail, parole, and sentencing. Moreover, because AI is data driven, some argue that a well-designed algorithm could in theory be more neutral or objective than a human. While AI invariably contains bias, a particular application might, in theory, be less biased than a human subject to implicit or explicit bias. All of these assumptions can be contested, in the abstract as well as with reference to specific AI applications.

Opponents of predictive algorithms make the following arguments.

Western law and criminal procedure are premised on individualized suspicion. This means an individual should be investigated or prosecuted based on articulable facts about them, not patterns found in data about the past conduct of other persons who may simply share one or more social descriptors, or data about past police practices and prosecutorial decisions. This argument is also especially applicable in the sentencing context.

All algorithms are biased in some way by the choices their human designers make: what metrics are used to evaluate data to make predictions, what data the algorithm is trained on, and what data the algorithm is tested on. Further, algorithms reflect human bias and can multiply and magnify bias by repeating it at scale. One of the most common criticisms of criminal risk assessment tools, for example, is that they rely on historical records of arrests, charges, convictions, and sentences, though “[d]ecades of research have shown that, for the same conduct, African-American and Latinx people are more likely to be arrested, prosecuted, convicted, and sentenced to harsher punishments than their white counterparts.”¹⁶¹

Predictive algorithms focus on characteristics that are, at least purportedly, readily discerned and susceptible to data adaptation and recording. Classifications such as “race,” gender, marital status, family status, address, and education likely play a disproportionate role in algorithm design and operation. Conversely, in

160. *State v. Loomis: Wisconsin Supreme Court Requires Warning Before Use of Algorithmic Risk Assessments in Sentencing, Recent Case*, *supra* note 143, at 1535; Ellora Israni, *Algorithmic Due Process: Mistaken Accountability and Attribution in State v. Loomis*, JOLT Digest (Aug. 31, 2017), <https://perma.cc/9NDU-NHXX>.

161. Minow et al., *supra* note 159, at 1.

operation or design, algorithms are less likely to include subjective weights, like role models and community connections and participation, that might also predict behavior and perhaps do so more accurately.

Classifications used as factors in algorithmic predictions are subject to all the risk of bias, intended and unintended. Even when unintentional, this bias may infiltrate an application through training data, how computer engineers assign weights to factors, or the “learning” the AI does on the job.

Some factors do not account for variation or nuance. Factors that appear to be subject to yes/no answers, and thus data scoring, may in reality be more complex and fall along a continuum. “Race” and ethnicity are good examples, even if not used intentionally in predictive tools (one would hope), other than to counter historical or algorithmic bias. Even marital status, a seemingly objective data point, may fall on a contextual continuum ranging from stable to unstable, happy to unhappy. Depending on what an algorithm is intended to predict, nuance can make all the difference in outcomes.

All of these factors are compounded where there is an inability to understand or challenge the underlying algorithm. Judges, of course, will have to determine when algorithm transparency is required as a matter of law, including due process. Lack of transparency undermines the ability of judges and litigators to assess the accuracy and meaning of an algorithmic output by asking questions like:

- What factors, including those that might serve as proxies for suspect categories, did the algorithm rely on?
- How were the factors weighted?
- Do those factors in fact reflect the case and parties in question?

We would encourage courts to look under the hood. Accurately understanding AI outputs often requires knowing not only the AI inputs but also the following: the weights that were attached and allocated to each input; the data the algorithm was trained, tested, and validated on; and an explanation of the methodologies used for prediction. Courts should ask:

- For what purpose was the AI designed, trained, tested, and validated? Is that the purpose for which the court is considering its use? If not, why is the court admitting or using the AI for an alternative purpose? (If, for example, a risk assessment was built to predict flight risk or future behavior, why is it being used at sentencing?) What safeguards is the court using or imposing to ensure appropriate use in context, and will the safeguards suffice?
- On what data was the AI trained, tested, and validated, and were those datasets biased? (See the “Probing for Bias” questions above.)
- What parameters did the algorithm search or use and what weight was given to those parameters in the AI output? Are such parameters and weights discoverable? Do those parameters include “race,” gender, other suspect

categories, or their proxies, such as housing and employment status,¹⁶² as factors? If so, why, and do they pass ethical and constitutional review?

- Does the AI have equal or disparate error rates or impacts across different racial, gender, or other suspect categories? (See the “Probing for Bias” questions above.) Did the developer and user of the AI perform and retain risk assessments and disparate impact assessments before using the AI, or in monitoring the AI post-deployment?
- Does the AI have a mechanism to evaluate its accuracy on an ongoing basis, specifically one to identify false positive and false negative rates, along with the trends associated with each?
- Was the defendant’s data included in the training data for the algorithm in question? If so, is it possible to expunge that data?
- Is the AI transparent in its function, its underlying data and factors, and its methodology of weighting? If not, is a more transparent methodology available?
- Does the application rely on a neural network? If so, what is the risk that the system will rely on parameters that are unintended or unknown to the designers or operators? Is it possible to identify those potential parameters?
- If the answers to these questions implicate proprietary information, what would prevent the court from hearing answers to the questions in camera, with appropriate protective orders and the power of contempt to enforce the court’s orders? Do the rights to due process, equal protection, and confrontation of witnesses require that the defendant or the public have access to all of the information at issue?

If the moving party cannot answer these questions to the satisfaction of the court, or is not prepared to answer these questions, judges might well be skeptical about the reliability of the evidence proffered or the use of a judicial tool.

Algorithms for Lawyers and Non-Lawyers: Technology-Assisted Review and . . . Legal Practice?

A variety of technologies offer litigation, drafting, and other tools for lawyers and non-lawyers. These technologies include: online legal advice tools; e-discovery tools; legal research; document management and creation; document- and case-level analytics; case outcome prediction (for forum and judge shopping); and

162. See Barabas et al., *supra* note 88, at 3.

lawyer-client matching.¹⁶³ We anticipate others. As with behavioral risk assessments, not all are AI-based, but an increasing number rely and will rely on machine learning and other AI methodologies.¹⁶⁴

Technology-assisted review (TAR), sometimes referred to as computer-assisted review (CAR) or “predictive coding,” is used for document review and production.¹⁶⁵ (TAR is not a brand name but rather a collective acronym for this type of software.) With the first generation of TAR (referred to as TAR 1.0), a lawyer first reviews and labels a set of “seed” documents for issues such as privilege, relevancy, and responsiveness, and then the algorithm learns to do the same based on the human-inputted labels.¹⁶⁶ The second generation (TAR 2.0) uses a similar model but with continuous active learning by the machine.¹⁶⁷ “Lawyers can also use TAR to structure the case, identify parties to depose, develop strategic defenses, and [complete] other tasks related to the case.”¹⁶⁸

Other examples of legal tools include those that provide automated legal advice, assist unrepresented litigants with legal proceedings, or draft documents, such as answers, discovery requests, motions, and simple briefs.¹⁶⁹

The quality of any such tools will depend significantly on the datasets available for training them; where data is not publicly available, large law firms and other businesses may have better access to confidential data (with permission from clients), and therefore better tools at their disposal.¹⁷⁰ This type of technology, however, does allow for the possibility of increasing access to justice for those who cannot afford traditional representation.¹⁷¹ It may also raise ethical issues about confidentiality, representation, competence, and diligence, as discussed earlier.¹⁷²

Legal research tools also use machine learning, for example to do natural language searches instead of Boolean searches,¹⁷³ or “analyze a draft argument to

163. David F. Engstrom & Jonah B. Gelbach, *Legal Tech, Civil Procedure, and the Future of Adversarialism*, 169 U. Pa. L. R. 1001, 1010–12 (2021).

164. *See id.* at 1015.

165. *Unlocking the E-discovery TAR Blackbox*, 102 Judicature No. 2, 63 (2018), <https://perma.cc/F3CS-9LM3>.

166. Kamron Sanders, *What Is Technology-Assisted Review? (TAR)*, Nat’l L. Rev. (May 19, 2022), <https://perma.cc/BQ4H-W7UP>.

167. *Id.*; Opentext, *Choosing the Right Technology-Assisted Review Protocol to Meet Objectives*, <https://perma.cc/4UBY-Z2NJ>.

168. Sanders, *supra* note 166.

169. Engstrom & Gelbach, *supra* note 163, at 1012.

170. *Id.* at 1018.

171. *Id.* at 1013, 1039.

172. *See id.* at 1013.

173. Boolean searches use words such as “and,” “or,” and “not,” and punctuation such as quotation marks to expand, limit, and qualify searches. Shauntee Burns, *What is Boolean Search?*, New York Public Library (Feb. 2011), <https://perma.cc/CL8E-4WP3>. The 19th century mathematician, George Boole, like many AI designers today, was seeking to use mathematical principles and logic

gain further insights or identify relevant authority that may have been missed.”¹⁷⁴ Many of these tools use natural language processing (NLP):

At a high level of abstraction, NLP aims to identify patterns in human language in ways that facilitate problem-solving. . . .

. . . The current research frontier, and a rapidly advancing one, is a mix of linguistics and “deep learning” (i.e., neural network) techniques. In a nutshell, deep-learning NLP machines make language computationally tractable by converting words, sentences, documents, or, in the legal context, entire cases into unique vectors, called “embeddings.” Each vector can be envisioned as an arrow from the origin to a point that represents the item of interest in a large, n-dimensional space, its magnitude a function of the presence of words, case citations, indexing concepts, or other features. Once this vast vector space has been constructed and human-annotated labels affixed to training materials (again, words, sentences, documents, cases), a sophisticated machine learning model can manipulate the vectors mathematically using large numbers (on the order of billions) of calculations to model relationships between them. With sufficient data and computing power, the system’s outputs enable a range of legal tasks, such as identifying relevant or privilege documents, past legal decisions that may be controlling, or, though . . . it is far trickier, the winning argument in a case.¹⁷⁵

A more complex and forward-leaning NLP application is Open AI’s ChatGPT 3 research app, publicly released in late 2022, and as of 2024, in a fourth public and for fee iteration (ChatGPT 4 and ChatGPT 4-Plus). Chat GPT 5 was released in summer 2025, showing the rapidity with which generative AI is developing. ChatGPT and other products use large language models (LLMs) to generate language or text, “writing” sentences that are often difficult to distinguish from a human’s. Open AI, a nonprofit funded by Microsoft, opened up ChatGPT to the public for use, in the process raising questions about work product and academic integrity across disciplines.¹⁷⁶ As described above, ChatGPT can incorporate research into its products as well, but that research may not always be reliable. The further development of this type of AI (large language models) presents fundamental questions for the legal profession—about the thoroughness and transparency of research and work product, about the very writing of law—that go well beyond the scope of this reference guide.¹⁷⁷ That said, many

to derive greater meaning from data and do so more rapidly than with existing 19th century constructs.

174. Matthew Stepka, *Law Bots: How AI Is Reshaping the Legal Profession*, ABA Bus. L. Today (Feb. 21, 2022), <https://perma.cc/54JD-KUTG>.

175. Engstrom & Gelbach, *supra* note 163, at 1020–21.

176. *Tools such as ChatGPT Threaten Transparent Science; Here Are Our Ground Rules for Their Use*, Nature (Jan. 24, 2023), <https://doi.org/10.1038/d41586-023-00191-1>.

177. Engstrom and Gelbach’s article, *supra* note 163, takes on many of these and other issues.

of the issues raised here, such as bias of all types and explainability and transparency, will shape future discussion.

AI and Discovery

If not in the form of an AI-generated brief (or hopefully and better, a human-generated brief aided perhaps by AI tools), a judge's consideration of AI at trial may start with discovery. Pretrial discovery is governed by the Constitution, case law, Federal Rule of Civil Procedure 26 (among other civil rules), and Federal Rule of Criminal Procedure 16. Rule 26 mandates parties make certain initial disclosures, "without awaiting a discovery request," including "a copy—or a description by category and location—of all documents, electronically stored information, and tangible things that the disclosing party has in its possession, custody, or control and may use to support its claims or defenses; unless the use would be solely for impeachment."¹⁷⁸ Additionally, with respect to expert testimony, "a party must disclose to the other parties the identity of any witness it may use at trial to present evidence under Federal Rule of Evidence 702, 703, or 705."¹⁷⁹ Unless otherwise stipulated or ordered by the court, the disclosure must be accompanied by a written report providing "a complete statement of all opinions the witness will express" and "the facts or data considered by the witness in forming them." Rule 16 of the Federal Rules of Criminal Procedure states, *inter alia*, that upon request "the government must permit the defendant to inspect and to copy or photograph. . . documents [and] data. . . within the government's possession, custody, or control [if] (i) the item is material to preparing the defense; (ii) the government intends to use the item in its case-in-chief at trial; or (iii) the item was obtained from or belongs to the defendant."¹⁸⁰ Documents and data might be argued to include documentation of algorithms and design and the data used to train, test, and validate AI models.

We forecast that as part of the discovery process at least four issues relating to AI will arise, and arise with frequency:

Timing, timeliness, and timelines: As noted above, Rule 26 provides for certain initial disclosures regarding documents and potential expert witness testimony, "without awaiting a discovery request." "A party must make the initial disclosures at or within 14 days after the parties' Rule 26(f) conference unless a different time is set by stipulation or court order." We surmise that, in some contexts, 14 days may be an unrealistic period of time for a party to determine which data and how much to turn over where the party intends to rely on AI-generated outputs as

178. Fed. R. Civ. P. 26.

179. Fed. R. Civ. P. 26(a)(2). Federal Rules of Evidence 702, 703, and 705 address "testimony by expert," "bases of an expert," and "disclosing the facts or data underlying an expert," respectively.

180. Fed. R. Crim. P. 16(a)(1)(E).

evidence. Consider a medical malpractice suit placing at issue a doctor's use of AI imagery as a diagnostic tool. Litigators are not likely to voluntarily turn over data without first being ordered to do so following a discovery conference.

Interpretation and Scope of Data Discovery: The timeline will also be impacted by the uncertain scope of Rule 26 when it comes to how much of the underlying data is required to be disclosed where an expert witness relies on an AI output “or electronically stored data that may be used to support its claims of defenses.” Judges, for example, will surely need to resolve on a case-by-case basis the parameters of “data considered by the witness in forming” an opinion when that opinion derives from the use of AI or AI outputs. Referring to our medical malpractice suit, for example, the parties are likely to dispute whether “data considered in forming an opinion” should include the data, or a characterization of the data, used to train, test, and validate an AI screening algorithm. An AI specialist could well argue that such data is necessary to determine whether the algorithm was trained to detect the disease in question, or to detect the disease in question with a comparative population and demographic base.

Similar interpretive issues are likely to arise in a criminal discovery context where, for example, the government intends to introduce AI outputs, such as a potential facial or voice recognition match placing a defendant at the scene of a crime. The challenge for courts and for discovery is that most facial or voice recognition tools are predictive rather than determinative in nature, and most are more accurate when more potential matches are identified (assuming that a true match is in the database). Restated, there are few Jason Bourne moments where the target is conclusively identified. The question for criminal discovery is just how much discovery a court should allow into the underlying AI tools based on the argument that any information—the algorithmic math, the training data, testing data, and validating process, for example—might undercut the predictive accuracy of the recognition tool in the case at hand and thus serve as potentially exculpatory evidence.

Supplementing Disclosures and Responses: Federal Rule of Civil Procedure 26 states that parties must supplement or correct their disclosure or response “in a timely manner if the party learns that in some material respect the disclosure or response is incomplete or incorrect.” AI tools are often iterative and changing, like a generative AI that is absorbing and incorporating new data into its outputs. That may present two challenges, one on either side of the same coin. First, in what manner to delimit the duty to supplement where the ongoing use of an AI tool may reveal earlier strengths and weaknesses at the time the output evidence was generated. Here the qualifier “material” will carry considerable weight and likely draw judges into disputes over whether iterative AI changes are material to understanding AI outputs offered into evidence. Second, the iterative nature of AI may make it difficult to fix and assess the quality and accuracy of AI outputs in a moment in time, such as the time and date an AI tool was used to assess a medical condition or may have impacted the performance of an AI-empowered vehicle.

The duty of continuing disclosure in Federal Rule of Criminal Procedure 16 is drafted differently than Rule 26. Rule 16(1)(c) states that: “A party who discovers additional evidence or material before or during trial must promptly disclose its existence to the other party or the court if: (1) the evidence or material is subject to discovery or inspection under this rule; and (2) the other party previously requested, or the court ordered, its production.” Where the government relies on an AI-generated output as evidence, updated information associated with that AI might be “additional new evidence or material.” Where new information about an AI output in a criminal context might be viewed as exculpatory under *Brady*,¹⁸¹ it might be constitutionally required as well, as in the case of new information decreasing the probability that an AI facial recognition algorithm correctly identified the defendant as a suspect. Similarly, if an AI model itself were to be considered a government witness, disclosures of material evidence that might impeach its veracity or accuracy might be required under *Giglio*.¹⁸²

Protective Orders, and Proprietary and Trade Secrets: One issue we forecast that will be argued in court, but for which courts are already well empowered and equipped, is the protection of trade secrets embedded in AI in the form of algorithms, data, parameters, and weights. Some litigants using AI in court, such as the government in the case of *Loomis*,¹⁸³ have argued that they cannot disclose the underlying algorithms, parameters, and weights, even to the court, because the underlying information is proprietary. We are skeptical of such assertions, which may arise in a variety of civil and criminal contexts. Civil Rule 26 provides clear and broad authority for courts in civil litigation to issue protective orders “requiring that a trade secret or other confidential research, development, or commercial information not be revealed or revealed only in a specified way.” Likewise, Criminal Rule 16 provides courts authority to issue protective and modifying orders “at any time the court may, for good cause, deny, restrict, or defer discovery or inspection, or grant other appropriate relief,” and do so *ex parte*. As discussed above, in the specific context of courts themselves using predictive algorithms for bail, probation, sentencing, or similar purposes, there are weighty due process, equal protection, and Confrontation Clause issues that might require *public* disclosure of algorithms and data, and indeed suggest any use of such tools is unconstitutional. In other contexts, it would seem the moving party should either be prepared to disclose material information about the AI outputs they intend to use at trial (potentially subject to appropriate protective orders) or the court should decline to admit such outputs into evidence. Where there is room for reasonable

181. *Brady v. Maryland*, 373 U.S. 83, 87 (1963); *Giglio v. United States*, 405 U.S. 150, 154 (1972).

182. See *Giglio*, *supra* note 181, at 154.

183. *Loomis v. Wisconsin*, 881 N.W.2d 749, 759 (Wis. 2016), *cert. denied*, 137 S. Ct. 2290 (2017).

debate, and judicial discretion and ruling, is on whether a particular parameter, dataset, or algorithm, for example, warrants protection or is in the public domain.

Judges as AI Gatekeepers

As evidentiary gatekeepers, judges will need to determine whether and when AI evidence will assist the factfinder and is admissible in court. The Federal Rules of Evidence and their state equivalents will help guide this determination. The Supreme Court’s *Daubert*,¹⁸⁴ *Crawford*,¹⁸⁵ and *Carpenter*¹⁸⁶ cases may also inform the evidentiary questions presented by AI. Neither these cases nor the Rules, however, were written with AI in mind. And currently few federal or state cases or jury instructions address AI. Judges will, of course, interpret and apply these cases and rules to AI in the specific contexts presented and do so consistent with the law of the jurisdiction in which they preside.

We briefly discuss here how the Federal Rules of Evidence and the *Daubert* factors might apply to AI-generated evidence not as a legal primer for an audience that needs no such explanation, but to highlight the complexity of applying the rules of evidence and case law to AI. We identify four threshold considerations when considering AI evidence given its technological characteristics and iterative nature. Courts, in time, will use their knowledge of the law to address myriad constitutional and evidentiary questions AI may raise.

Federal Rules of Evidence 401–403, 702, 901–902

As judges know, under Federal Rule of Evidence 401, evidence is relevant if “(a) it has any tendency to make a fact more or less probable than it would be without the evidence; and (b) the fact is of consequence in determining the action.”¹⁸⁷ Rule 402 states that relevant evidence is admissible unless the Constitution, a federal statute, the other Federal Rules of Evidence, or other rules prescribed by the Supreme Court apply and would exclude the evidence.¹⁸⁸ Due process or Confrontation Clause concerns, for example, might bar or limit certain AI evidence from admission. Statutes addressing data privacy and use may do so as well. Rule 403 allows a court to exclude relevant evidence if its probative value is substantially outweighed by a danger of: creating unfair prejudice; confusing

184. *Daubert v. Merrell Dow Pharms., Inc.*, 509 U.S. 579 (1993).

185. *Crawford v. Washington*, 541 U.S. 36 (2004).

186. *Carpenter v. United States*, 138 S. Ct. 2206 (2018).

187. Fed. R. Evid. 401.

188. Fed. R. Evid. 402.

the issues; misleading the jury; causing undue delay; wasting time; or needlessly presenting cumulative evidence.¹⁸⁹

Many of the threshold evidentiary issues associated with AI will be litigated under Rules 402 and 403 or their state equivalents. Relevancy in most or all jurisdictions is broadly defined, and most AI applications are essentially tools for assessing probability, in theory, making them inherently relevant in assessing, under Rule 401, whether something is “more or less probable.” (Litigants might still argue that certain AI outputs are too inaccurate or biased to be relevant or that they are inapt for the purpose offered.) The primary issues, then, are (1) the reliability of AI generally and (2) the appropriateness of use in the context presented. Rules 402 and 403 are pertinent because the evidentiary use of AI will invariably present questions about discovery and due process, such as whether there is a right to access an underlying algorithm or data used to generate evidence or inform judicial decisions. Another issue is the risk that litigation over AI will present the figurative “trial within a trial” and potentially confuse the jury under Rule 403. Also, courts might apply Rule 403 to exclude AI evidence that is biased or otherwise unreliable. Inquiry is prudent; otherwise, juries may assume AI evidence has the imprimatur of “science” or “technology” in the context presented, potentially lending it false authority or undue weight, or permitting its use in a manner for which it was not intended.¹⁹⁰

Judges will need to decide in what manner and to what extent to require authentication of the AI evidence offered and how, if at all, to validate its reliability. These criteria will likely bring Federal Rules of Evidence 702 and 902 into play, as well as *Daubert* and *Crawford*.

AI and the interpretation of AI outputs are complex. Courts will have to determine the appropriate means to verify AI outputs. This might involve expert testimony, or it might be done through technical means, such as cryptographic hashes embedded in an image at the time it is created. Courts will need to determine who is qualified to testify about the accuracy and fairness of an AI application. Of course, steady, purposeful, and consistent application of the Federal Rules of Evidence or their state equivalents on the record is a good place to start.

It is intuitive, but worth remembering that proponents of AI-generated evidence will seek to simplify its admission by limiting or eliminating as many threshold foundational requirements as possible. Opponents of admission will seek to undermine its relevance and reliability in general or for the purpose for

189. Fed. R. Evid. 403.

190. Andrea Roth, *Machine Testimony*, 126 Yale L. J. 1972 (2017) (“Moreover, just as the Framers were concerned that factfinders would be unduly impressed by affidavits’ trappings of formality, ‘computer[s] can package data in a very enticing manner.’ The socially constructed authority of instruments, bordering on fetishism at various points in history, should raise the same concerns raised about affidavits.”) (internal citations omitted).

which it is offered. To challenge relevance and accuracy, opponents will seek access to the underlying algorithm, the data on which it was trained, tested, and validated, as well as knowledge of what occurs and what is weighted inside any machine-learning black box. Thus, courts will face a layered adjudicative challenge each time AI-generated evidence is offered.

Where AI outputs are admitted, opponents may seek to cross-examine the software engineers responsible for its design. Each AI application is different. It will have a different purpose, rely on a different algorithm, use a different machine learning methodology or methodologies, and will train, test, and validate using different data. Consequently, AI issues are generally not subject to resolution through the application of case-law precedent in the same way that, for example, DNA analysis is now widely accepted in court. A single or lead case will not likely resolve when to admit AI-generated evidence. Adjudication is to be expected for each application and in each context for which the application is offered as evidence. Further, because AI is a constellation of technologies and applications, it is not a single process or technology that can be validated once and generally adopted. In each instance where AI evidence is offered, there may be a legitimate need to explore the underlying technology and different components of that technology for use in that instance or for the proffered purpose. Courts should also consider the following analytical factors.

Context

Courts should pay attention to whether a particular AI application is a good “fit”¹⁹¹ for the purpose for which it is proffered. Some criminal risk assessment algorithms, for example, are designed for the purpose of determining which individuals might benefit from alternatives to incarceration, such as parole or counseling. These algorithms might have less relevance and reliability if used to determine sentencing.¹⁹² That will depend on all the factors noted above, including the input factors, the weight assigned to those factors, the data on which the algorithm was trained, tested, and validated, and the nature of the confidence thresholds applied to the output. Courts should pause and ask not only whether the AI at issue is relevant and material to the matter before the court but for what purpose the AI was specifically designed and whether the outputs will materially and fairly inform the fact finder. Courts should inquire about the contents of the datasets on which the algorithm was trained, tested, and validated.

191. See *Daubert*, 509 U.S. at 591–92.

192. See Christopher Bavitz et al., *supra* note 158, at 6–7 (discussing the Wisconsin Supreme Court’s warning in *State v. Loomis*, 881 N.W.2d 729 (Wis. 2016) that the risk assessment tool COMPAS was not developed for use at sentencing).

Case-Specific Reliability

Even when an algorithm is being used for the purpose for which it was designed, there may be data or design reasons why output reliability will decrease in a specific context. An AI algorithm may have been designed for and tested on a population substantially different from the population for which the output is offered, with less accurate results than the lab-tested confidence threshold¹⁹³ (“inappropriate deployment” as discussed above in the section titled “AI Is Biased”). As previously noted, according to the FBI, the Bureau’s facial recognition application has an 86% percent match rate (confidence threshold) when an input image is compared to at least fifty potential output matches drawn from state license databases. However, the same algorithm would not have the same match accuracy if run against a different input demographic, say, the population of another country—not because the algorithm is necessarily intentionally biased but because it has not been trained against a comparative population pool. (In fact, output disparity across gender and ethnicity has been an issue with some facial recognition algorithms.¹⁹⁴ As a result, facial recognition accuracy has been a focal point of AI design initiatives, and therefore we anticipate future American facial recognition applications will help mitigate this issue.)

Inapt Factors

There is a risk with ML that a neural network will rely on inapt factors in making its output predictions. Judges will want to know whether this is possible and, if so, regarding which factors, before allowing a jury to assess the weight of AI evidence or before judges use an algorithm themselves to assess bail or recidivism risk. For example, a judge would want to determine, consistent with case law and the Constitution, which factors were included and weighted within any AI-driven bail, parole, confinement, or sentencing tool, to ensure that inappropriate, inapt, or unconstitutional factors were not included and that any factors, even if appropriate, were not given undue weight by the neural network. A judge would also want to know if any factors might be working as proxies¹⁹⁵ for suspect categories.

193. See Barabas et al., *supra* note 88, at 3, and Bavitz et al., *supra* note 158, at 7.

194. NIST, *NIST Study Evaluates Effects of Race, Age, Sex on Face Recognition Software: Demographics study on face recognition algorithms could help improve future tools* (Dec. 19, 2019), <https://perma.cc/NKN2-PLDA>.

195. See Barabas et al., *supra* note 88.

Underlying Bias

Courts will also want to investigate the ways in which a given AI application is biased, or may be biased, before admitting its outputs into evidence or relying on it to inform a judicial decision, as discussed earlier.

Daubert

One way to conceptualize AI and appreciate its evidentiary complexity in relation to other technology is to apply the (non-exhaustive) list of factors the Supreme Court developed in *Daubert v. Merrell Dow Pharmaceuticals, Inc.*¹⁹⁶ to determine whether expert testimony based on a specific, scientific methodology should be admitted.¹⁹⁷ *Daubert* and, in certain states, its predecessor, *Frye v. United States*,¹⁹⁸ govern the admission of expert testimony based on scientific methodology. *Daubert* uses a “factors” approach, while *Frye* uses a “general acceptance” standard. Each of the *Daubert* factors opens wide the door to debate over many AI attributes. Analysis under *Frye*, too, is likely to be complicated, requiring judges to determine when a scientific method is “sufficiently established to have gained general acceptance in the particular field to which it belongs.”¹⁹⁹ In theory, making such a determination will entail examining not only the specific algorithm and use in question but also identifying the relevant field of acceptance and what acceptance means for something like facial recognition or behavior prediction—all in a context where algorithms may be iterative and/or the data they are trained on are iterative and changing.

The *Daubert* factors include:²⁰⁰

- whether the theory or technique in question can be and has been tested,
- whether it has been subjected to peer review and publication,
- its known or potential error rate,
- the existence and maintenance of standards controlling its operation, and
- whether it has attracted widespread acceptance within a relevant scientific community.

With AI, these factors would need to be applied to individual algorithms and applications rather than “AI” generally, which, as stated above, generically

196. *Daubert*, 509 U.S. at 579.

197. We are not aware of a court having done so at the time this reference guide was drafted.

198. *Frye v. United States*, 293 F. 1013 (D.C. Cir. 1923).

199. *Id.* at 1014.

200. *Daubert*, 509 U.S. at 593–95.

describes a constellation of technologies and methodologies. We address each factor in turn.

Testing

The first step suggested by *Daubert* is to identify the theory, technique, or component that is subject to evaluation. There are many options with AI. Is it: The sensor(s) that fed data to the AI system? The algorithm? The math behind the algorithm? The dataset used to train the algorithm? The training methodology? Or is it the system as an integrated whole that is subject to review?

The second step is to decide what test is appropriate and what baseline to use to establish accuracy. Medical diagnostic AI, for example, might be compared to physician-diagnosed outcomes. It is true that medical diagnostics are subject to social influence and human and machine bias. But in medicine there is often a fixed data point, an established fact or yes/no answer to whether a disease or tumor is present, against which testers can measure the algorithm's accuracy.

In contrast, an algorithm intended to predict future behavior, such as a criminal assessment tool, cannot be tested with the same degree of scientific or evidence-based meaning, given the weight placed on social factors. Recidivism algorithms, for example, attempt to predict future human behavior, using circumstantial factors drawn from a base population. In such contexts, there is no certain result and no control group, and confirming predictions is difficult. Human circumstances are endlessly complex, creating multiple influences on behavior—without necessarily *determining* behavior. Nor is there a way to verify, after an individual has been jailed or sentenced, how an individual's future behavior is affected by imprisonment. The experience of imprisonment itself might turn a person toward or away from future crime, making it difficult or impossible to verify the machine's prediction. In short, predictive algorithms in the criminal context are especially difficult to test, to peer review, and to assess for accuracy and error rates.

Peer Review

An example innovation in AI-enabled medicine highlights the question of machine reliability and illustrates the importance of peer review. In April 2019, NPR reported that Stanford computer scientists had created an algorithm for reading chest X-rays to diagnose tuberculosis.²⁰¹ They hoped to use it to diagnose the disease in HIV patients in South Africa, and the machine's results were

201. Richard Harris, *How Can Doctors Be Sure A Self-Taught Computer Is Making The Right Diagnosis?*, NPR (Apr. 1, 2019), <https://perma.cc/Y2WM-FF4N>.

already better than doctors’.²⁰² To corroborate their success, the Stanford scientists submitted their results to other scientists for review.²⁰³ One noticed a peculiarity in the AI’s decision making.

[The peer reviewers] Zech and his medical school colleagues discovered that the Stanford algorithm to diagnose disease from X-rays sometimes “cheated.” Instead of just scoring the image for medically important details, it considered other elements of the scan, including information from around the edge of the image that showed the type of machine that took the X-ray. When the algorithm noticed that a portable X-ray machine had been used, it boosted its score toward a finding of TB.

Zech realized that portable X-ray machines used in hospital rooms were much more likely to find pneumonia compared with those used in doctors’ offices. That’s hardly surprising, considering that pneumonia is more common among hospitalized people than among people who are able to visit their doctor’s office.

“It was being a good machine-learning model and it was aggressively using all available information baked into the image to make its recommendations,” Zech says. But that shortcut wasn’t actually identifying signs of lung disease, as its inventors intended.²⁰⁴

The machine was making a correlational, rather than causal, connection between the use of a portable machine and TB. Without informal peer review, humans might not have discovered that aspect of how the AI algorithm was making decisions, a clear example of both how AI adapts and often does so in the black box. The TB-scan example also demonstrates the disruptive role of unknowns, here an unwitting, algorithmic bias (“inappropriate focus”). The original programmers evidently did not anticipate that the machine would teach itself to evaluate information beyond the scan itself. It is impossible for a programmer to anticipate every real-world factor a machine will encounter and attempt to interpret.

Peer review may be particularly challenging with AI. For example, it is hard to imagine a credible peer review that does not include access to the underlying algorithm, parameter weights, and the data on which an algorithm was trained, tested, and validated. How else could a peer validate the accuracy and use of an application? Moreover, peer review should be case or use specific. An AI that is good at predicting the need for substance abuse counseling may not be suited for predicting other types of risk or need. The usefulness of any peer review will also vary in accordance with the talent and capability of the individuals or entities conducting the peer review.

202. *Id.*

203. *Id.*

204. *Id.*

Error Rates

Judges will also need to ask the right questions to determine whether error rates are accurate and meaningful. For example, will or might error rates vary depending on whether the AI application is tested and reviewed using the relevant local population (database) to which it will be applied, as opposed to a national population, or perhaps a more idealized lab database?²⁰⁵ What types of bias might be affecting the accuracy of any reported error rates? (See suggested questions under “Probing for Bias,” above)

Standards Controlling an AI Application’s Operation and Maintenance

AI imposes operational and maintenance obligations. At this point in time, however, operational standards, if any, are set voluntarily. The Intelligence Community and the Department of Defense have each published principles for the ethical use of AI, while many companies have their own internal standards. In the absence of uniform statutory standards, courts might begin by asking: What dataset is used? Is that dataset updated appropriately? Is the machine learning monitored by continued testing against known results to ensure the machine is not learning bad habits? Courts might also ask all the questions about bias suggested in “Probing for Bias,” above.

Acceptance

Courts will also need to determine what widespread acceptance within the relevant scientific community means in the context of AI. There is a big difference between general acceptance of the field and acceptance of a specific application. Many computer engineers and government actors accept the premise and use of facial recognition, but privacy advocates do not. Skepticism will remain with any specific application. The point is also illustrated by driverless cars. General acceptance of the concept has not at present translated to acceptance of a model of fully autonomous driverless car that is ready for commercial sale and public use. What then would constitute appropriate general acceptance?

Proprietary Algorithms

How does one test the accuracy or conduct a peer review of a proprietary algorithm or an iterative or evolving ML algorithm? Google is not likely to disclose

205. See Barabas et al., *supra* note 88, at 3, and Bavitz et al., *supra* note 158, at 7.

its search algorithm for public or peer inspection and risk its market position in the search engine arena. Unless courts can demonstrably protect such trade secrets while also testing their validity, applying the *Daubert* factors to many or most AI applications in open court may be difficult. (Jurists or lawmakers²⁰⁶ may determine defendants or the public should have access to certain underlying algorithms and data, such as in instances where liberty interests are at stake. Courts will need to determine whether the Fifth, Sixth, and Fourteenth Amendments require it.)

In other contexts, where courts seek to allow litigants to test the validity of AI applications while still protecting proprietary information, they might exercise their general power to oversee how evidence is entered, to enforce rulings, and to seal records. A parallel can be found in the way classified information is protected while still allowing certain litigation to proceed, with records reviewed by judges and sometimes by cleared counsel. Also relevant is the 1996 Defend Trade Secrets Act (18 U.S.C. § 1835), which specifically directs federal courts to protect trade secrets in proceedings arising under Title 18 of the U.S. Code. Section 1835(a) states,

In any prosecution or other proceeding under this chapter, the court shall enter such orders and take such other action as may be necessary and appropriate to preserve the confidentiality of trade secrets, consistent with the requirements of the Federal Rules of Criminal and Civil Procedure, the Federal Rules of Evidence, and all other applicable laws.

In context, specific statutes also provide intellectual property protections for AI, such as those protections found in § 705 of the Defense Production Act (DPA), which allow the president in the first instance and courts in the second instance, through the power of contempt and jurisdiction found in § 706, to protect intellectual property relevant to DPA enforcement or defend against DPA actions.

Deepfakes and Evidence

AI's capacity to convert symbolic language (coded numbers) into natural language and to discern, recognize, and formulate patterns at the pixel level makes it a tool of choice not only for identifying voices and pictures but also for mimicking voices and altering images. Moreover, AI can do so with real-life

206. For example, an Idaho law, Section 19–1910 of the Idaho Code, states, “All pretrial risk assessment algorithms shall be transparent, and all documents, records, and information used to build or validate the risk assessment shall be open to public inspection, auditing, and test. No builder or user of a pretrial risk assessment algorithm may assert trade secret or other protections in order to quash discovery in a criminal matter by a party to a criminal case.” <https://perma.cc/J3QQ-GTZ4>.

precision, creating images or recordings known as “deepfakes.” Hollywood has, of course, known about deepfakes for years, though in movies they’re called “special effects,” as in *Star Wars* or *Forrest Gump*. What makes deepfakes noteworthy for courts is not only the lifelike quality attainable but the accessibility of this capability to the general population. Tools now readily available on the internet allow non-specialists to alter photographs and mimic speech with startling realism, capable of fooling practically everyone—including triers of fact.

Deepfakes today are often created with “generative adversarial networks,” or GANs. GANs “are two-part AI models consisting of a generator that creates samples [e.g., of images, film, or audio] and a discriminator that attempts to differentiate between the generated sample and real-world samples.”²⁰⁷ The discriminator (or sometimes multiple discriminators) provides feedback to the generator that helps it improve its realism.²⁰⁸

Deepfakes can have artistic and educational value, such as in bringing historical figures back to life in film, but also nefarious purposes, such as the creation of non-consensual pornography and national security threats.²⁰⁹ Deepfake technology, its “accessibility [by the public], usability, and the verisimilitude of its outputs,” will only improve over time; meanwhile, the U.S. government, academia, nonprofits, and industry all have sought to improve technology for detecting deepfakes,²¹⁰ creating a bit of an arms race.

In addition to improving deepfake detection, technologists have also sought to improve methods for authenticating genuine records.²¹¹ There are tools and methods to authenticate images, such as using a cryptographic hash, a numerical value based on the zeros and ones in the digital sequence of an image.²¹² For example, *The Economist* reports of an app, eyeWitness to Atrocities, developed by a London charity, that:

When a photo or video is taken by a phone fitted with that app, it records the time and location of the event, as reported by hard-to-deny electronic

207. Kyle Wiggers, *Generative Adversarial Networks: What GANs Are and How They’ve Evolved*, VENTUREBEAT (Dec. 26, 2019), <https://perma.cc/LLR4-G5LP>; see also Riana Pfefferkorn, “Deepfakes” in the Courtroom, 29 B.U. Pub. Int. L.J. 245, 249 (2020), <https://perma.cc/9T7R-3GAU>.

208. Wiggers, *supra* note 207.

209. Danielle K. Citron & Robert Chesney, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, 107 Cal. L. Rev. 1753 (2019); Rebecca A. Delfino, *Pornographic Deepfakes: The Case for Federal Criminalization of Revenge Porn’s Next Tragic Act*, 88 Fordham L. Rev. 887 (2019).

210. Pfefferkorn, *supra* note 207, at 250.

211. *Id.* at 251

212. See NIST Special Publication 800-152, *A Profile for U.S. Federal Cryptographic Key Management Systems* (Oct. 2015), <https://perma.cc/45Q6-K3ZQ>. This defines “hash function” as “[a]n algorithm that computes a numerical value (called the hash value) on a data file or electronic message that is used to represent that file or message, and depends on the entire contents of the file or message. A hash function can be considered to be a fingerprint of the file or message.”

witnesses such as GPS satellites and nearby mobile-phone towers and Wi-Fi networks. This is known as the controlled capture of metadata, and is more secure than collecting such metadata from the phone itself, because a phone's time and location settings can be changed. Second, the app reads the image's entire digital sequence (the zeros and ones which represent it) and uses a standard mathematical formula to calculate an alphanumeric value, known as a hash, unique to that picture.²¹³

The metadata is sent to a secure server, separate from where the image is stored. Later, the image can be examined, and any difference from the recorded hash value would indicate an alteration in the image.²¹⁴ Such a process depends on calculating and recording the numeric value at the time of origin in a chain of custody manner. In short, there are different and developing mechanisms to tag social media posts, pictures, and other data. As discussed below, one question is when courts should require factual as well as legal authentication before admitting images, videos, or voice recordings into evidence.²¹⁵

For all these reasons, deepfakes may present evidentiary challenges to civil and criminal justice systems—for example, if the video or audio recording is a deepfake used to show guilt or innocence, or liability or lack thereof. A litigant might also proffer a deepfake as evidence without realizing it is a deepfake.²¹⁶ Scholars have also expressed concerns that a jury's general awareness of deepfakes, or a litigant's false defense that something is a deepfake, may also cause a jury to discount a genuine recording and generally undermine faith in the justice system.²¹⁷

Under Federal Rule of Evidence 901, “To satisfy the requirement of authenticating or identifying an item of evidence, the proponent must produce evidence sufficient to support a finding that the item is what the proponent claims it is,” with the jury or trier-of-fact ultimately determining the weight of the evidence.²¹⁸ Scholars identify some risks, including:

213. See *Proving a Photo Is Fake Is One Thing. Proving It Isn't Is Another*, *The Economist* (Jan. 9, 2023), <https://perma.cc/2F8T-JMER>.

214. *Id.*

215. Media outlets and NGOs like Bellingcat have developed the use of AI applications that can help identify the location of a photo or video as well as the date and time it was taken. There are also AI apps that reverse search photos to match those photos with historical or social media photos to verify locations and personal identities. There are AI apps as well that can colorize black and white photos to help match them against colored photos. Such tools have proven useful in documenting and validating the documentation of war crimes committed by Russian soldiers and units in Ukraine. See Foeke Postma, *Using New Tech to Investigate Old Photographs*, Bellingcat (Aug. 9, 2022), <https://perma.cc/T7E5-G89N>.

216. Pfefferkorn, *supra* note 207, at 255.

217. *Id.*; see also Rebecca Delfino, *Deepfakes on Trial: A Call to Expand the Trial Judge's Gatekeeping Role to Protect Legal Proceedings from Technological Fakery*, 74 *Hastings L. J.* 293 (2023), <https://perma.cc/5URU-HBUA>.

218. John P. LaMonaga, *A Break From Reality: Modernizing Authentication Standards for Digital Video Evidence in the Era of Deepfakes*, 69 *Am. U. L.R.* 1945, 1963 (citing *United States v. Branch*,

Reference Guide on Artificial Intelligence

- as deepfake technology becomes better and more accessible, an authenticating witness may not be able to tell whether an item proffered as evidence is genuine or fake; nor would jurors be able to do so later, when weighing the evidence;
- a witness might allow a deepfake to inform and shade imperfect human memory; and
- where authentication is based on the fact that an item comes from an archive or system of records (for example, police body cam video archives), possible manipulation via cyber-hacking or otherwise might not be evident.²¹⁹

Courts and legislatures might examine whether existing rules of evidence for authentication and expert witness testimony are sufficient to address the issue of deepfakes. Scholars diverge on this point.²²⁰ Scholars and practitioners tend to agree, however, that trials will require expert testimony by digital forensic specialists as well as increased litigation over the authenticity and veracity of photographic, video, and audio evidence.²²¹

Federal Rule of Evidence 902(13) and (14), adopted in 2017, set forth “certified records generated by an electronic process or system” and “certified data copied from an electronic device, storage medium, or file,” respectively, as “items of evidence that are self-authenticating.” Items with cryptographic hashes, discussed above, might be authenticated under these rules. The committee notes on the 2017 amendments suggest that 902(13) and (14) were adopted to avoid “the expense and inconvenience of producing a witness to authenticate an item of electronic evidence” only to have the other party stipulate to authenticity or fail to challenge the authentication testimony. The amendments do, however, provide for a party to challenge self-authentication: “the proponent must give an adverse party reasonable written notice of intent to offer the record—and must make the record and certification available for inspection.”²²² A party suspicious of a deepfake would have “fair opportunity to challenge” the record and certification.²²³ Again, digital forensic experts would be in demand.

970 F.2d 1368, 1370 (4th Cir. 1992)); Naciye Celebi, Qingzhong Liu, & Muhammed Karatoprak, *A Survey of Deep Fake Detection For Trial Courts*, <https://perma.cc/42BY-QEKJ>.

219. Pfefferkorn, *supra* note 207, at 250; *see also* Delfino, *supra* note 217, at 330–35.

220. *See* Pfefferkorn, *supra* note 207, at 259–75 (rules suffice), and Delfino, *supra* note 217, at 332–48 (suggesting changes).

221. *Id.*

222. Fed. R. Evid. 902(11), which is incorporated in the quoted part by reference in Fed. R. Evid. 902(13) and (14).

223. *Id.*

Conclusion

The multidisciplinary AI field presents a myriad of complex evidentiary challenges. The law rarely, if ever, keeps pace with technology. The legislative and appellate processes simply do not move at the same pace as technological change and could not if they tried. Moore's Law is faster than case law. Likewise, scholars and commentators are currently better at asking questions than they are at answering them. Artificial Intelligence itself is a fast-moving field encompassing a constellation of technologies.

Judges and lawyers do not need to be mathematicians or coders to understand AI and to wisely adjudicate the use of AI in courts or by courts. Judges need to define and understand their roles as overseers of discovery, evidentiary gatekeepers, constitutional guardians, translators, and in some cases, potential AI consumers. Perhaps the most important thing courts can do is ask careful and informed questions. Judges should also put their analysis and application of the answers on record as to whether, why, how, and subject to what evidentiary standards and determinations AI has been admitted into evidence or used by courts. This will allow full and informed appellate review. We hope this reference guide provides a foundation judges can use for understanding AI and building a common law of AI.

Glossary of Terms

This glossary is borrowed from our publication *An Introduction to Artificial Intelligence for Federal Judges*,²²⁴ co-authored by Matthew Mittelsteadt.

algorithm. “[A] step-by-step procedure for solving a problem or accomplishing some end.”²²⁵ A familiar example is a recipe, which details the steps needed to prepare a dish. In a computer, an algorithm is implemented in computer code and details the discrete steps and calculations a computer needs to implement to complete a task. An algorithm is the “engine” an AI uses to “think” and make predictions. In the field of AI, the term “algorithm” is often used synonymously with “computer program.” A program, however, is a more specific term, referring to an algorithm written in computer code and packaged for execution.

algorithmic bias. According to McKinsey, “[w]hile ‘bias’ can refer to any form of preference, fair or unfair,” undesirable AI bias is bias that leads to “discrimination against certain individuals or groups of individuals based on the inappropriate use of certain traits or characteristics.”²²⁶ As AI is designed by humans, AI will always be biased and will assume the biases of its engineers, potentially leading to poor or discriminatory results. As noted throughout this guide, causes of bias can stem from the statistical, such as errors in model design, to the social, to the use of inapt data to the context presented. Bias can also be caused by incomplete datasets. For instance, a facial recognition AI trained only on male faces may perform poorly when analyzing female faces.

artificial general intelligence (AGI). In the future it is possible we will move past narrow AI and develop artificial general intelligence that does not have a narrow function and can serve multiple purposes. AGI can be conceived of as an AI system that equates to or surpasses the general-purpose intelligence of the human brain.²²⁷ AGI does not have a precise definition, and the line that divides it from narrow AI is grey. As such, the introduction of AGI will likely be a gradual process, and it is unlikely there will be a precise “Sputnik moment” that introduces the age of AGI.

artificial intelligence (AI). There is no agreed upon or general definition of AI. However, one practical definition is that artificial intelligence is any machine

224. James Baker, Laurie Hobart, & Matthew Mittelsteadt, *An Introduction to Artificial Intelligence for Federal Judges* (Federal Judicial Center 2023), <https://perma.cc/7DJT-T9D2>.

225. Merriam-Webster, *Definition of “algorithm”* (May 20, 2021), <https://perma.cc/82FG-EJRD>.

226. Jake Silberg, *Notes from the AI Frontier: Tackling Bias in AI (and in Humans)*, McKinsey Global Institute (June 2019), <https://perma.cc/K8Z8-WUAE>.

227. *What Is Artificial Intelligence (AI)?*, IBM, *supra* note 68.

that can “perform tasks that would otherwise require human intelligence.”²²⁸ The NSCAI expands on this idea, noting “AI is not a single piece of hardware or software, but rather, a constellation of technologies that gives a computer system the ability to solve problems and to perform tasks that would otherwise require human intelligence.”²²⁹ AI is implemented in computers as algorithms based on models such as artificial neural networks (ANN), which are often iteratively designed through methods such as machine learning (ML) or deep learning (DL). One statutory definition of AI is found in the National Defense Authorization Act of 2020, which states:

The term “artificial intelligence” means a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations or decisions influencing real or virtual environments. Artificial intelligence systems use machine and human-based inputs to—

- (A) perceive real and virtual environments;
- (B) abstract such perceptions into models through analysis in an automated manner; and
- (C) use model inference to formulate options for information or action.

Section 5002(3) of the NAI Act 2020 (Division E of NDAA 2020).

artificial neural network (ANN). The model (or “tool”) used in deep-learning AI best defined as a computer system that works to achieve intelligence through a network structure that works to simulate the human brain.²³⁰ An ANN analyzes data by passing data through multiple layers of artificial neurons that sift through and decipher the data. This layered network structure allows the system to analyze discrete data elements, draw connections between discovered data patterns, and ultimately derive meaning and form predictions. Neural networks can be wide, meaning each layer has large numbers of neurons, or deep, meaning data must pass through many layers of neurons before a final conclusion is drawn. Engineers determine the width and depth of the network based on their interpretation of the tools and structures a specific AI application needs for success.

autonomous systems. AI-controlled machines and vehicles such as driverless cars and aerial drones that can operate and make decisions with little or no human control. Such systems already exist. In most cases, stringent safety demands have forestalled widespread use.

228. Baker, *supra* note 37, at 21.

229. NSCAI, *supra* note 1, at 8.

230. *What Is a Neural Network?*, IBM, *supra* note 110.

black box. A term used to describe the often-mysterious nature of AI decision making and the problem of AI explainability.²³¹ Most AI programs write their own algorithms through machine learning, which can result in complex code and decision-making processes indecipherable even to engineers. Such complexity limits human ability to understand how an AI makes decisions, and what factors, including biases, may have influenced those decisions. However, considerable research is underway by organizations such as the National Institute of Standards and Technology (NIST) to enable more transparent neural networks, which may allow judges and lawyers to more fully understand the parameters and weights applied within.²³²

confidence score. Any expression of certainty in the predictive accuracy of an AI or ML application.²³³ AI applications are imperfect and offer approximate results, decisions, or predictions that can be provided with a level of confidence. Few, if any, results an AI produce should be treated as a certainty. For example, the FBI facial identification software mentioned in the introduction to this guide is not designed or intended to match a single identity with a face. Rather it offers the user a range of potential matches based on potential pattern similarities or matches. The algorithm is reported to be accurate 86% of the time when the algorithm output offers the user at least fifty potential match pictures.²³⁴ Put another way, the AI has 86% confidence that the match will be one of the fifty given matches.

deep learning. A machine-learning approach characterized by the use of a multilayered artificial neural network.²³⁵ Deep learning can use but does not necessarily require labeled datasets.²³⁶ This approach has exploded in popularity over the last decade²³⁷ and is the predominant form of machine-learning AI. Common applications use deep learning, including many driverless cars and voice-recognition AI.

facial recognition. A prominent class of AI applications that can detect a face and analyze its features (or “biometrics”) and even predict the identity of that face. These AI applications are notable for their common use in criminal justice and national security as a means of identifying suspects or threats.

231. Ariel Bleicher, *Demystifying the Black Box That Is AI*, Scientific American (Aug. 9, 2017), <https://perma.cc/5BFP-WRW6>.

232. NIST, AI Fundamental Research—Explainability (June 16, 2022), <https://perma.cc/3YX5-JC8Q>.

233. Jason Brownlee, *Confidence Intervals for Machine Learning*, Machine Learning Mastery (Aug. 8, 2019), <https://perma.cc/VGT2-282N>.

234. GAO, *supra* note 2, at 14.

235. *What Is Machine Learning (ML)?*, IBM, *supra* note 111.

236. *Id.*

237. *What Is Artificial Intelligence (AI)?*, IBM, *supra* note 68.

Facial recognition algorithms can also be used to surveil more generally. Facial recognition may also be used as a biological “password” to authenticate an individual’s identity (for example, to unlock a smartphone).

human in-the-loop. An autonomous AI system designed to work cooperatively with a human to complete its tasks. Often these AI defer to human judgment when making certain decisions, especially those with significant consequences or moral weight. Human in-the-loop systems generally seek a “best of both worlds” approach that maximizes the benefits of both human and AI decision making.

human on-the-loop. An autonomous AI system designed to work under human oversight, allowing the human to easily intervene if the AI’s decisions are in error, pose significant danger, or are ethically compromising.

human out-of-the-loop. An autonomous AI system designed to operate without human oversight or involvement. Such systems do not facilitate easy human intervention if unethical or dangerous decisions are made.

lethal autonomous weapons systems (LAWS or simply AWS). Autonomous systems that can use deadly force, which have received outsized legal, ethical, and political attention given widespread concerns about giving inhuman systems the power to take a human life.

machine learning (ML). A method of creating AI that relies on data, algorithms, and learned experience to refine algorithms and form intelligence.²³⁸ The premise of machine learning is that “intelligence” is not innate but must be learned through experience. Machine-learning AI algorithms are “trained” by engineers who feed it mass amounts of data that it slowly learns to interpret and understand. In response to the data, the AI gradually tweaks its code to steadily improve its abilities. These tweaks add up over time, helping the AI create stronger predictions. Forms of machine learning include the following:

- **supervised learning** (“learning through instruction”): A form of machine learning where engineers specify a desired outcome and feed the AI algorithm curated and labeled data to guide AI toward that outcome.²³⁹ For example, to teach a facial recognition AI to match names and faces, labeled facial data would be fed to the AI algorithm so it could learn which faces correspond to which names. This method is ideal for tasks with agreed upon “correct” answers or decisions.
- **unsupervised learning** (“self-taught learning”): A form of machine learning where unstructured and uncured data are fed to a machine-learning algorithm that finds trends, patterns, and relationships in that

238. *What Is Machine Learning (ML)?*, IBM, *supra* note 111.

239. *Id.*

data.²⁴⁰ This is useful for finding insights humans may have overlooked or cannot perceive. This method is ideal for applications without a firm answer and for general data analysis.

- **reinforcement learning** (“learning by doing”): A form of machine learning where the algorithm learns through trial and error.²⁴¹ Its learning is often guided by a goal (e.g., winning a game of chess) and adjusts its parameters to better reflect trials that helped it to reach or come close to that goal. This method is useful for discovering optimal solutions in many rule-based systems, including chess, chemistry, physics, and traffic pattern analysis.

narrow AI. “[T]he ability of computational machines to perform singular tasks at optimal levels, or near optimal levels, and usually better than, although sometimes just in different ways, than humans.”²⁴² Under this umbrella falls many single-purpose or limited-purpose AI technologies, such as facial recognition algorithms, driverless cars, and drones. These technologies are intelligent in one or a few domains, limiting their ability to handle complexity or tasks outside of their intended purpose. All AI currently in use falls in this category.

natural language processing (NLP). AI algorithms designed to process, analyze, and recognize written or verbal human speech at human levels.²⁴³ NLP has a wide variety of applications. Familiar applications include virtual assistants such as Amazon’s Alexa or Apple’s Siri. In national security and criminal justice, NLP can be used to analyze and understand language recordings and written information, drawing conclusions, insights, and patterns from that data. This offers a powerful intelligence and investigative tool. It can also be used to match a voice to an identity (much like facial recognition) and for language translation.

superintelligence (SI). AI philosophers also contemplate the emergence of superintelligence, a stage of AI evolution marked as beyond human intelligence.²⁴⁴ SI has sparked widespread concern, and its risks and benefits are unclear. Stephen Hawking highlighted this uncertainty in 2015, exclaiming that “AI may be the best thing to ever happen to humanity or the worst.”²⁴⁵ A malicious SI could cause incalculable harm and a beneficial SI could prove an invaluable tool. It must be noted that many engineers and government officials

240. *Id.*

241. *Id.*

242. Baker, *supra* note 37, at 29.

243. *What Is Natural Language Processing (NLP)?*, IBM, <https://perma.cc/23HX-3JPN>.

244. *What Is Artificial Superintelligence?* IBM, <https://perma.cc/K834-Z3CW>.

245. Alex Hern, *Stephen Hawking: AI Will Be “Either Best or Worst Thing” for Humanity*, *The Guardian* (Oct. 19, 2016), <https://perma.cc/2DGX-25HF>.

disdain consideration of superintelligence as the stuff of science fiction as well as a distraction from the real and immediate challenges of narrow AI today.

Acknowledgments

The views expressed in this reference guide are our own, as are any errors. However, we would like to thank the following people for their help: Matthew Mittelsteadt, for helping to develop and edit the ideas herein; our research assistants over several years, Rebecca Buchanan, Thomas Clifford, Shannon Cox, Henry DuBeau, Aaron Ernst, Thomas Finnigan III, Hannah Gabbard, Rickson Galvez, Harrison Gregoire, Kaitlyn Keane, Alyssa Kozma, R.J. Naperkowski, Carlos Negron, Margaret Santandreu, Michael Stoianoff, and David Trombly, for their hard work and good humor; Kristen Duda, for all her help in managing our team; the Federal Judicial Center and National Academies of Sciences, Engineering, and Medicine, including Jason A. Cantone, Joe Cecil, Nathan Dotson, José Idler, Steven Kendall, Dominic LoBuglio, Anne-Marie Mazza, and Beth Wiggins, for the opportunity to publish this reference guide and their review and other work on the reference guide; the Hon. Curtis Collier, who provided helpful feedback as part of the FJC review process; and reviewers and editors of a shorter version of this guide (published by the Georgetown Center for Security and Emerging Technology), Chuck Babington, Keith Bybee, Tobias Gibson, Danny Hague, Matt Mahoney, Hon. John Sparks, and Lynne Weil. We also thank the Committee on Science for Judges and our NAS reviewers, known to us as A, B, and C, for their helpful comments, and Geoff Erwin for his careful and detailed editing. Finally, we thank our families and loved ones who gave us the time, patience, and support we needed to immerse ourselves in AI and draft this reference guide. Thank you all.

Reference Guide on Climate Science^{*}

JESSICA WENTZ AND RADLEY HORTON

Jessica Wentz, LL.M., is a Non-Resident Senior Fellow at the Sabin Center for Climate Change Law, Columbia Law School.

Radley Horton, Ph.D., is a Professor at the Columbia Climate School, Columbia University.

CONTENTS

Introduction, 1563

Foundational Components of Climate Science, 1565

Scope of Research, 1565

Core Concepts and Methods, 1566

Physical Understanding, 1566

The greenhouse effect and radiative forcing, 1568

Biogeochemical cycles, 1569

Atmospheric and ocean circulation and the hydrological cycle, 1570

Feedback loops, tipping points, and cascading impacts in the climate system, 1571

Natural variability, 1573

Climate Datasets, 1574

Statistical Techniques and Climate Models, 1575

Managing and Communicating Uncertainty, 1578

Sources of Climate Research, 1580

Scientific and Consensus Reports, 1580

Peer-Reviewed Research, 1582

Expert Testimony and Reports, 1583

Climate Change Detection, Attribution, and Projections, 1585

Detection and Attribution, 1586

General Methods and Parameters, 1588

Detection of change, 1588

Attribution to anthropogenic climate change, 1588

Challenges associated with downscaling and exogenous variables, 1590

^{*}Following the December 31, 2025, publication of the Fourth Edition of *Reference Manual on Scientific Evidence*, on February 6, 2026, the Federal Judicial Center decided to omit the climate science chapter from their website.

Climate Change Detection and Attribution, 1592	
Methods and parameters, 1592	
Status of research, 1594	
Extreme Event Detection and Attribution, 1599	
Methods and parameters, 1599	
Status of research, 1603	
Impact Detection and Attribution, 1608	
Methods and parameters, 1609	
Status of research, 1610	
Source Attribution, 1615	
Methods and Parameters, 1615	
Status of Research, 1619	
National contributions, 1619	
Corporate contributions, 1622	
Projections of Future Climate Change, 1623	
Methods and Parameters, 1623	
Status of Research, 1625	
How Climate Science Factors into Litigation, 1627	
Overview of Climate Litigation, 1628	
Legal Applications of Climate Science, 1631	
Causation and Harm, 1631	
Foreseeability, 1637	
Legal Obligations and Authorities, 1638	
Acknowledgments, 1640	
Glossary of Terms, 1641	
References, 1645	
Select References on Climate Science, 1645	
Select References on Climate Science and the Law, 1645	
Appendix: Overview of Scientific Organizations and Government Agencies Involved in the Production, Synthesis, and Dissemination of Climate Science, 1646	
World Meteorological Organization (WMO), 1646	
Intergovernmental Panel on Climate Change (IPCC), 1646	
Other Major Scientific Organizations, 1649	
Government Agencies, 1650	
Nongovernmental Organizations, 1652	

FIGURES

1. Changes in global surface temperature relative to 1850–1900, 1596
2. Emission scenarios in IPCC AR6, 1626

Introduction

Climate science examines the structure and dynamics of the Earth's climate system and how that system is influenced by human and natural processes.¹ The climate system consists of five interacting systems: the atmosphere, hydrosphere, cryosphere, lithosphere, and biosphere.² Thus, climate science spans multiple disciplines, including atmospheric science, physical geography, and oceanography. It also encompasses research on the interactions between the global climate system and other natural and human systems, which implicates fields such as biology, economics, and social sciences. Owing to the complexity and pervasiveness of the climate system, it is difficult to establish precise boundaries for this field—but for the purposes of this reference guide, we focus on the physical science aspects of climate science.³

Climate scientists use observational data, physical understanding, statistical analysis, and climate models to understand and characterize the complex interactions among different components of the climate system. One key focus of this research is to understand the mechanisms of observed climate change and to characterize how humans are affecting the global climate system through greenhouse gas (GHG) emissions and other climate forcing factors.⁴ The Intergovernmental Panel on Climate Change (IPCC), an intergovernmental organization tasked with assessing scientific knowledge about climate change, is generally viewed as the leading scientific authority in this field.⁵ The IPCC periodically publishes assessment reports synthesizing the latest research on climate change. The IPCC's Sixth Assessment Report (AR6) found that: “[i]t is unequivocal that human influence has warmed the atmosphere, ocean and land,” and this has caused “[w]idespread changes in the atmosphere, ocean, cryosphere, and

1. For a more detailed definition, see *Climate Science*, in Stanford Encyclopedia of Philosophy (May 11, 2018), <https://perma.cc/SCN3-5Y4M>.

2. “Climate” describes weather conditions averaged over a period of time, typically decades or more. The atmosphere is the gaseous envelope surrounding the Earth; the hydrosphere refers to the components of the earth system composed of water; the cryosphere refers to the components of the earth system composed of ice; the lithosphere is the rocky outer portion of the Earth; and the biosphere refers to all life on Earth. Please refer to the glossary for more detailed definitions of these and other technical terms used in this chapter.

3. Although our focus is on physical climate science, it is necessary to discuss the interactions between physical climate science and other fields when discussing certain areas of the science. See, e.g., sections titled “Impact Detection and Attribution” and “Source Attribution” below.

4. A climate forcing factor is any substance, activity, or event that affects the flow of energy coming into or out of the global climate system, thus affecting the amount of heat retained within the system. Anthropogenic climate forcers include GHGs, aerosols, and changes in land use that make land reflect more or less solar energy. There are also natural climate forcers, such as solar radiation and the particulate matter from volcanic eruptions.

5. For more information about the IPCC and its assessment methodology, see section titled “Scientific and Consensus Reports” and Appendix below.

biosphere.”⁶ The global warming attributable to human activities is unprecedented in the last 2,000 years⁷ and “is already affecting every inhabited region across the globe, with human influence contributing to many observed changes in weather and climate extremes.”⁸

The increasing effect of anthropogenic climate change has contributed to growth in climate-change-related litigation in the United States and other jurisdictions. Much of this litigation seeks to establish legal obligations on the part of governments to control GHG emissions, prepare for the effects of climate change, and disclose climate-change-related risks in government documents; some litigation seeks the opposite. There are several ways in which climate science may factor into the resolution of climate-related lawsuits. First, climate science can be used to assess causation—for example, whether and to what extent an actor has caused or contributed to climate-change-related risks or injuries. Second, climate science can be used to assess whether climate-related impacts or risks are foreseeable, which is potentially relevant to determining whether a defendant is required or authorized to take some action in response to climate change. Third, climate science can be used to determine the scope of a defendant’s legal obligations and authorities—for example, whether the defendant has an obligation to reduce GHG emissions or take measures to prepare for the impacts of climate change.

The purpose of this reference guide is to help judges evaluate the admissibility and weight of expert testimony and documentary evidence involving climate science.

The second part of this reference guide, “Foundational Components of Climate Science,” describes the data and methodologies used in climate research, the scientific disciplines that constitute this field, the types of qualifications held by experts in this field, and the primary sources of climate research.

The third part of this reference guide, “Climate Change Detection, Attribution, and Projections,” provides a more in-depth discussion of research on anthropogenic climate change. It describes research methodologies and the status of scientific knowledge across three disciplines: (1) climate change detection and attribution research, which examines whether observed changes in natural and human systems can be attributed to human influence on climate; (2) source attribution research, which examines the relative contributions of different entities to

6. IPCC, *Climate Change 2021: The Physical Science Basis*, Contribution of Working Group I to the Sixth Assessment Report of the IPCC (Valérie Masson-Delmotte et al. eds.) [hereinafter IPCC AR6 WGI], <https://perma.cc/D93G-EPRC>. Other scientific bodies have also found unequivocal evidence of anthropogenic climate change. See National Academy of Sciences, *Climate Change: Evidence and Causes: Update 2020* [hereinafter NAS Update], <https://perma.cc/9XTW-S7DA>; U.S. Global Change Research Program, *Fourth National Climate Assessment*, Vol. 1, *Climate Science Special Report* (Donald J. Wuebbles et al. eds., 2017) [hereinafter NCA4 Vol. I], <https://perma.cc/EQ2X-AMEP>.

7. IPCC AR6 WGI, *supra* note 6.

8. *Id.* at 10.

anthropogenic climate change; and (3) predictive research, which provides insights on future climate change and its impacts under different emissions trajectories and warming scenarios.

The last part of this reference guide, “How Climate Science Factors into Litigation,” discusses the role of climate science in litigation. It provides a brief overview of the types of legal claims that implicate climate science, focusing on federal claims, and describes how different areas of climate research may factor into judicial assessments of causation, foreseeability, and legal obligations and authorities. There are various contexts where judges may confront genuine scientific questions and areas of uncertainty in this field—for example, when tasked with determining whether a specific plaintiff’s injuries were caused by climate change, and whether and to what extent a defendant contributed to those injuries. This reference guide seeks to provide the scientific context necessary for answering such questions, while also recognizing the normative and legal considerations that will factor into this type of analysis.

Foundational Components of Climate Science

Scope of Research

The Earth’s climate system is enormous and complex, consisting of many nested and interlinked subsystems, including the atmosphere, hydrosphere, cryosphere, lithosphere, and biosphere. The interactions between these components produce the conditions known as “climate”—i.e., weather conditions averaged over a period of time, typically decades or longer. Although the climate system is global, climate is typically characterized in reference to a particular region or locale as a result of spatial variations in climatological conditions.

Climate science is the field of study aimed at characterizing the climate system and understanding the physical processes and parameters that ultimately determine climate conditions. Owing to the breadth and complexity of the climate system, climate research spans many different scientific disciplines. For example, research on the physical science of climate and climate change (i.e., understanding of the physical properties of the climate system and how it is changing) encompasses: (1) mathematics and statistics; (2) basic sciences, e.g., physics and chemistry; (3) earth sciences, e.g., atmospheric sciences, climatology, physical geography, oceanography, meteorology, hydrology, biogeochemistry, and cryospheric sciences; and (4) computer sciences and data analysis. Most climate research involves collaboration across these different fields.

Climate science also encompasses research on the interactions between the climate and other natural and human systems, with a significant body of work

aimed at characterizing the effects of climate change on people, infrastructure, and ecosystems. Research on climate change impacts implicates both physical climate science (including the subfields noted above) and other disciplines, such as (1) biological and ecological sciences; (2) social sciences and economics; (3) epidemiology, population health, and health impact research; (4) fire dynamics; and (5) engineering and other fields relevant to assessing impacts to transportation, energy, and other infrastructure systems. Although studies on climate impacts vary considerably in terms of their focus (i.e., the impact(s) being studied), these studies typically utilize common methodologies, datasets, and models, as discussed below.

Core Concepts and Methods

The foundational components of climate science are:

- **Physical understanding of the climate system**, i.e., understanding of the physical properties of the climate, including the different component systems and interactions between those systems, as well as the effect of exogenous variables on the system.
- **Climate datasets**, which include direct observations of climate variables, paleoclimate reconstructions, and reanalysis datasets.
- **Statistical techniques** and **climate models**, which are used to evaluate patterns, trends, causal relationships, variability, and uncertainty within the climate system, and to develop projections of future climate change.

These components are interconnected and mutually reinforcing—for example, physical understanding of the climate system is based on observations and statistical analysis, and climate models incorporate a combination of physical understanding, observational evidence, and statistical techniques in order to derive insights on how changes in a particular input to the climate system can affect other variables within the system. The sections below provide more detailed explanations of each component and its role in climate research.

Physical Understanding

In the early 1800s, physicists and other scientists began conducting experiments and developing theories to explain the physical mechanisms behind global and regional climate conditions. It was hypothesized that the Earth had undergone significant climate changes in the past based on observations of the natural

world, but the mechanism behind these changes had yet to be discovered.⁹ In the 1820s, the physicist Joseph Fourier developed an early theory of what would eventually be recognized as the “greenhouse effect.” Fourier had calculated that an object that is the same size as the Earth and same distance from the sun would be considerably colder than the Earth if it were heated by incoming solar radiation alone. Fourier hypothesized that the Earth’s atmosphere trapped incoming solar radiation in the form of heat, thus causing the planet to be warmer than would otherwise be the case.¹⁰ Several decades later, scientists such as Eunice Foote and John Tyndall recognized that atmospheric gases, particularly carbon dioxide (CO₂) and water vapor, may be responsible for this warming effect. They correctly hypothesized that these gases absorbed incoming solar radiation—preventing it from being re-emitted to space—thus increasing the energy and heat content of the Earth’s atmosphere. These scientists were also able to demonstrate and measure the greenhouse effect of CO₂ through lab experiments in the 1850s and 1860s.¹¹ Subsequently, in the late nineteenth century, the Swedish chemist Svante Arrhenius hypothesized that human-caused GHG emissions from fossil fuel use and other combustion sources were large enough to cause global warming.¹²

Since those early experiments, scientific knowledge of the climate has advanced significantly. IPCC AR6 found that physical understanding of the fundamental features of the climate system is “robust and well established,”¹³ and scientists can explain many of the core processes that determine climate conditions. Some areas of physical understanding that are integral to climate research include: (1) the greenhouse effect and radiative forcing; (2) the carbon cycle and other biogeochemical cycles that govern the movement of chemical elements among different parts of the climate system and connected systems; (3) the mechanisms of atmospheric and ocean circulation and their relationship to the hydrological cycle; (4) the role of feedback loops, tipping points, and cascading impacts in the climate system; and (5) natural variability in the system.

9. For example, in the early 1800s, Jean-Pierre Perraudin hypothesized that glaciers might be responsible for giant boulders seen in alpine valleys, and Luis Agassiz subsequently hypothesized that glaciers had covered much of Europe and North America during what he referred to as an “Ice Age.” See E. P. Evans, *The Authorship of the Glacial Theory*, 145 N. Am. Rev. 94 (1887), <https://perma.cc/Y9AG-4P9P>.

10. Thomas R. Anderson et al., *CO₂, The Greenhouse Effect and Global Warming: From the Pioneering Work of Arrhenius and Callendar to Today’s Earth System Models*, 40 *Endeavour* 178 (2016), <https://doi.org/10.1016/j.endeavour.2016.07.002>.

11. W.F. Barrett, *Contributions to Molecular Physics in the Domain of Radiant Heat*, 7 *Nature* 66 (1872), <https://doi.org/10.1038/007066a0>; Eunice N. Foote, *Circumstances Affecting the Heat of the Sun’s Rays*, 22 *Am. J. of Sci. & Arts* 382 (1856), <https://perma.cc/X6F9-UVT6>.

12. Svante Arrhenius, *On the Influence of Carbonic Acid in the Air upon the Temperature of the Ground*, 41 *London, Edinburgh, & Dublin Phil. Mag. & J. Sci.* 237 (1896), <https://perma.cc/MV2A-UCJT>.

13. IPCC AR6 WGI, *supra* note 6, at 44, 150.

The greenhouse effect and radiative forcing

The greenhouse effect plays a fundamental role in shaping climate conditions, as it ultimately determines the amount of incoming solar energy that is retained within the system. The heat-trapping properties of GHGs are now well understood and can be quantified over different time horizons. Water vapor is the most abundant naturally occurring GHG in the atmosphere and is responsible for approximately half of the Earth's natural greenhouse effect. Other common GHGs in the atmosphere include CO₂, methane (CH₄), nitrous oxide (N₂O), fluorinated gases, chlorofluorocarbons (CFCs), and hydrochlorofluorocarbons (HCFCs). As discussed in the second part of this reference guide, "Climate Change Detection, Attribution, and Projections," large increases in GHG concentrations due to fossil fuel emissions and other human activities are the dominant cause of observed global warming and climate change.¹⁴

The effects of anthropogenic GHG emissions on the atmosphere are an example of radiative forcing—a change in the energy flux within the Earth's atmosphere. Positive radiative forcing occurs when the Earth receives more incoming energy from sunlight than it radiates into space, and this net gain of energy causes warming. There are a number of natural processes that can affect net radiative forcing—these include changes in the percentage of incoming solar radiation absorbed by the Earth, volcanic activity, orbital cycles, and changes in global biogeochemical cycles (discussed below). There are also other human drivers that can affect atmospheric energy flux—for example, land use changes can have positive or negative effects on radiative forcing, and aerosol emissions have negative radiative forcing and thus contribute to cooling of the climate system.

Scientists have been systematically studying the interactions between both human and natural forcing since the early 20th century.¹⁵ Although the heat-trapping properties of specific substances are fairly well understood, there is still some uncertainty about the effect of certain processes (e.g., land use changes) on climate forcing, as well as the relative contributions of different forcing agents to observed climate changes, especially at regional scales. Nonetheless, physical understanding of radiative forcing associated with GHGs is sufficiently robust to

14. Observations have demonstrated that atmospheric moisture content (water vapor) is increasing along with other GHGs, and this amplifies the greenhouse effect. However, the increase in atmospheric moisture content is directly attributable to human-induced global warming (from GHG emissions). In other words: increased water vapor is a consequence of anthropogenic climate forcing; it is not an indicator of "natural" climate change. Were these other GHG concentrations to somehow drop to pre-industrial levels, water vapor would almost instantaneously return to pre-industrial levels as well. See B.D. Santer et al., *Identification of Human-Induced Changes in Atmospheric Moisture Content*, 104 Proc. Nat'l Acads. Scis. [hereinafter PNAS] 15248 (2007), <https://doi.org/10.1073/pnas.0702872104>.

15. IPCC AR6 WGI, *supra* note 6, at 150.

support assessments and projections of global warming. For example, each additional increment of CO₂ added over the past century and projected to be added for the coming decades is expected to yield a comparable change in direct radiative forcing. Furthermore, past projections of global temperature change in response to radiative forcing have been broadly consistent with subsequent observations.¹⁶

Biogeochemical cycles

Biogeochemical cycles govern the transfer of chemicals among different components of the climate system and other earth systems. Some of the cycles that are integral to the study of climate include the carbon cycle, nitrogen cycle, water cycle, and phosphorus cycle. Here we focus on the carbon cycle owing to the dominant role of CO₂ in global climate change. The water (or hydrological) cycle is also integral to the study of climate, as discussed in various contexts below.

The carbon cycle describes how carbon moves between the atmosphere, hydrosphere, biosphere, cryosphere, and lithosphere.¹⁷ Most of the carbon on the planet is stored in the lithosphere, primarily in sedimentary rock deposits, and a small fraction stored in fossil fuels. Carbon is also stored in the atmosphere (as CO₂),¹⁸ in the oceans (as dissolved atmospheric carbon and in carbonate sediments), and in the biosphere (as organic molecules in organisms and soil). Oceans and terrestrial systems are generally characterized as “sinks” or “reservoirs” of carbon because, at today’s high atmospheric concentrations of CO₂, they typically absorb more carbon than they release into the atmosphere.

The primary pathways through which carbon is released into the atmosphere include human combustion of fossil fuels, industrial processes, wildfires, volcanic eruptions, biological respiration, and decomposition of organic matter. Once CO₂ enters the atmosphere, it can remain there for a very long time—potentially thousands of years or more—but there is significant variation in the atmospheric lifetime of individual CO₂ molecules.¹⁹ A large proportion (20–30%) of the CO₂ that enters the atmosphere is absorbed by the oceans, where it dissolves into seawater and forms carbonic acid (resulting in ocean acidification). The initial process of ocean absorption can occur within a relatively short time frame (e.g., within five years), but much of that carbon is circulated back into the

16. *Id.* at 150.

17. For a more detailed overview of the carbon cycle and how it is influenced by human activities, see U.S. Global Change Research Program, *Second State of the Carbon Cycle Report*, (N. Cavallaro et al. eds., 2018), <https://perma.cc/KT2H-GTUK>.

18. Atmospheric carbon also comes from methane (CH₄), which is converted into CO₂ as it combines with oxygen.

19. Because of variation, the IPCC does not recognize a specific lifetime for CO₂ molecules. See IPCC AR6 WGI, *supra* note 6, at 302.

atmosphere on a short time frame as well (e.g., within ten years or less).²⁰ Thus, carbon is constantly cycling between the oceans and atmosphere.

Carbon is also constantly cycling between the atmosphere and biological systems on land (i.e., forests, grasslands, and other ecosystems). The amount of carbon stored in the terrestrial sinks is largely based on the carbon uptake of vegetation through photosynthesis, with large interannual variability due to natural changes in vegetation as well as human land uses. There are some instances in which terrestrial sinks may become carbon sources—specifically, when the amount of carbon that is released is greater than the amount of carbon that is absorbed. This may occur, for example, as a consequence of wildfires or ecosystem degradation.²¹

Human activities, including land use changes and the burning of fossil fuels, have disturbed the natural carbon cycle, releasing more than two trillion metric tons of CO₂ and methane (CH₄) into the atmosphere since the onset of the industrial revolution. A portion of those emissions remains in the atmosphere; the remainder is absorbed by ocean and terrestrial sinks. The “airborne” fraction of anthropogenic carbon emissions has remained constant at approximately 44% over the past six decades,²² but the capacity of ocean and terrestrial sinks to absorb carbon is expected to decline as the CO₂ concentrations increase.²³

Atmospheric and ocean circulation and the hydrological cycle

The movement of air, water, and heat in the climate system plays an integral role in shaping regional climate conditions, including regional temperature, precipitation, humidity, and aridity, as well as extreme weather events. In particular, regional climate is influenced by atmospheric circulation (the large-scale movement of air and heat in the atmosphere), ocean circulation (the large-scale movement of water in oceans), and the hydrological cycle (the circulation of water between different earth systems), in addition to other factors.²⁴ Understanding

20. Mason Inman, *Carbon Is Forever*, 1 *Nature Climate Change* 156 (2008), <https://doi.org/10.1038/climate.2008.122>.

21. See, e.g., Sirui Wang et al., *Potential Shift From a Carbon Sink to a Source in Amazonian Peatlands Under a Changing Climate*, 115 *PNAS* 12407 (2018), <https://doi.org/10.1073/pnas.1801317115>.

22. IPCC AR6 WG1, *supra* note 6, at 676.

23. *Id.* at 677.

24. Some of the other factors that shape regional climate include the amount of sunlight the area receives (which depends, in large part, on latitude); altitude, topographical features, and the shape of the land (i.e., “relief”); and proximity to oceans and large water bodies.

the mechanisms and relationships between these large-scale processes is key to understanding the climate system and climate change.²⁵

These circulatory processes are governed by thermodynamics as well as other physical processes (e.g., gravity, surface friction, planetary rotation).²⁶ Thermodynamic processes are those that involve the exchange of heat, work, temperature, and energy. There is strong theoretical understanding of how thermodynamics influence certain aspects of the climate system, which contributes to high confidence findings for certain trends and impacts, particularly those that are directly attributable to temperature changes.²⁷ Radiative forcing is an example of a thermodynamic process that is well understood. Dynamic processes are those that deal with the movement of bodies in response to physical forces (e.g., the physical transport of air masses of a given temperature and moisture content).²⁸ As noted by Screen et al. (2018), the dynamic manifestations of climate change are not as well understood as thermodynamic aspects, but they are strongly constrained by well-understood principles, especially the conservation of energy and mass, and the knowledge gap is narrowing as a result of recent research unpacking the causal mechanisms behind observed changes in circulation patterns.²⁹ More limited understanding of dynamic processes contributes to uncertainty about certain aspects of climate and climate change, including uncertainties about the relationships between increasing GHG concentrations, changes in hydrological and cryospheric processes, and changes in clouds.³⁰

Feedback loops, tipping points, and cascading impacts in the climate system

Understanding the climate system and anthropogenic influence on that system requires understanding of feedback loops, cascading impacts, and tipping points. Feedback loops are causal processes that occur when outputs of a system are routed back as inputs, resulting in acceleration of a process or change (positive

25. Coupled models have been developed to simulate the interactions between atmospheric and ocean circulation and other components of the climate system. See section titled “Statistical Techniques and Climate Models” below.

26. The hydrological cycle is also influenced by chemical and biological interactions, which is why it is characterized as a biogeochemical process.

27. IPCC AR6 WGI, *supra* note 6, at 430.

28. The term “dynamic” is also used in climate science to describe systems that are characterized by change and complexity. Here, we specifically refer to dynamics as a subdivision of physical mechanics.

29. J.A. Screen et al., *Polar Climate Change as Manifest in Atmospheric Circulation*, 4 Current Climate Change Reps. 383 (2018), <https://doi.org/10.1007/s40641-018-0111-4>.

30. See, e.g., Eilat Elbaum et al., *Uncertainty in Projected Changes in Precipitation Minus Evaporation: Dominant Role of Dynamic Circulation Changes and Weak Role for Thermodynamic Changes*, 49 Geophysical Resch. Letters e2022GL097725 (2022), <https://doi.org/10.1029/2022GL097725>.

feedback) or deceleration (negative feedback).³¹ One key example of a feedback loop in the climate system is water-vapor feedback: warmer temperatures increase water vapor in the atmosphere, and the water vapor, which is a GHG, traps additional heat. Researchers believe that this may play an important role in current and future warming trends.³² Another important feedback loop is the ice-albedo feedback loop, whereby warmer temperatures result in the melting of ice caps, glaciers, and sea ice (all of which have high albedo, i.e., they reflect more sunlight back to space), thus decreasing the albedo of land and ocean surfaces, and accelerating the warming process.

The term “tipping point” describes a threshold that, when surpassed, will result in large and typically irreversible changes.³³ Tipping points are common throughout the climate system as well as other systems that are affected by climate change. Key examples of important tipping points within the climate system are the temperature at which the melting of the Greenland ice sheet becomes irreversible (a process that would ultimately trigger meters of sea-level rise as well as changes in atmospheric and ocean dynamics), the collapse of Arctic winter sea ice, the dieback of the Amazon rainforest, the irreversible loss of mountain glaciers, and the collapse of boreal permafrost. Generally speaking, the existence of tipping points is known with much higher confidence than the temperature thresholds at which they will occur. Some critical tipping point thresholds may have already been surpassed, although the full effects have not yet manifested because of time lags and/or incomplete understanding.³⁴ This highlights an important aspect of tipping points: critical thresholds can be “locked in” before the actual event occurs (e.g., the near complete melting of the Greenland ice sheet may already be inevitable because of existing warming).³⁵ Although much is unknown about the timing and potential consequences of climate tipping points, there are significant risks associated with surpassing these thresholds, since consequences can be so large.³⁶

Cascading impacts are a related concept. These can be conceptualized as a series of impacts that occur together due to interdependencies within a

31. Dennis L. Hartmann, *Climate Sensitivity and Feedback Mechanisms*, in *Global Physical Climatology* (2d ed. 2016).

32. A.E. Dessler et al., *Stratospheric Water Vapor Feedback*, 110 PNAS 18087 (2013), <https://doi.org/10.1073/pnas.1310344110>.

33. The IPCC defines a tipping point as a “critical threshold beyond which a system reorganizes, often abruptly and/or irreversibly.” IPCC AR6 WG1, *supra* note 6, at 95.

34. David I. Armstrong McKay et al., *Exceeding 1.5° Global Warming Could Trigger Multiple Climate Tipping Points*, 377 Sci. 1171 (2022), <https://doi.org/10.1126/science.abn7950>.

35. Niklas Boers & Martin Rypdal, *Critical Slowing Down Suggests that the Western Greenland Ice Sheet Is Close to a Tipping Point*, 118 PNAS e2024192118 (2021), <https://doi.org/10.1073/pnas.2024192118> (finding that the Greenland ice sheet melt tipping point is between 0.8°C and 3.2°C of warming above pre-industrial levels).

36. Timothy M. Lenton et al., *Climate Tipping Points—Too Risky to Bet Against*, 575 Nature 592 (2019), <https://doi.org/10.1038/d41586-019-03595-0>.

system, flowing out to other domains, and potentially amplifying risks and hazards.³⁷ Cascading impacts can occur as a result of feedback loops and surpassing tipping points. Cascading impacts are a particular concern when evaluating the effects of climate change on human and natural systems. For example, changes in bioclimatic conditions caused by global warming can result in a cascade of ecosystem alterations (e.g., extinctions leading to alterations in food webs that have cascading effects on other species, potentially even triggering further extinctions).

Natural variability

Natural variability in the climate system refers to those variations in climate that are caused by events and processes of nonhuman origin, in conjunction with the chaotic nature of the system. It includes variability that is internally generated within the system, like the El Niño Southern Oscillation (ENSO),³⁸ as well as variability driven by natural external factors, such as variations in solar activity and volcanic eruptions. Natural variability can play a prominent role in explaining variations in global climate over certain time spans (years, months, days). Put another way, the chaotic nature of the climate system—whereby small differences in initial conditions can ultimately “excite” different patterns of internal variability at different points in time leading to large differences in climate states—means that even in the presence of anthropogenic forcing, a broad range of temperature trends ranging from positive to negative can be experienced in models—and presumably observations as well if we had more than one sample to draw from—on a timescale of years at the global scale, and decades at the regional scale. However, the influence of natural variability tends to be small when evaluating recent large-scale trends over periods of multiple decades or longer.³⁹ Understanding natural variability is important for climate change attribution and projections, insofar as it is necessary to ascertain whether an observed trend, impact, or event is a consequence of anthropogenic forcing rather than natural variability within the system. Some patterns of natural variability, like ENSO, may themselves be impacted by anthropogenic warming in ways that are not yet completely understood.

Although there is still some uncertainty about the physical drivers of natural climate variability, researchers have developed techniques for reproducing many aspects of variability in climate models based on increasing physical

37. See Judy Lawrence et al., *Cascading Climate Change Impacts and Implications*, 29 *Climate Risk Mgmt.* 1000234 (2020), <https://doi.org/10.1016/j.crm.2020.100234>.

38. ENSO is an interaction between the tropical Pacific atmosphere and ocean that produces roughly year-long global impacts approximately every five to seven years.

39. See IPCC AR6 WGI, *supra* note 6, at ch. 3.

understanding of the climate system.⁴⁰ The effects of natural variability are sometimes quantified using a signal-to-noise ratio, which compares the strength of an anthropogenic climate change signal against natural variability noise. The challenge of distinguishing the anthropogenic signal is more pronounced in regional climate assessments because natural variability plays a larger role in shaping regional (in contrast to global) climates (i.e., there is a larger signal-to-noise ratio at global scales, where some variability tends to cancel out). There can also be additional sources of internal variability at even finer scales, such as variations in ocean current location that can impact weather and climate at the scale of a nearby coastal city, for example. The accuracy with which scientists can distinguish anthropogenic signals from natural variability noise also depends on the type of impact being studied, and the temporal boundaries of the study.

Climate Datasets

In order to study climate and climate change, scientists need to be able to characterize past and present climate conditions across different geographic and temporal scales. They rely on climate datasets that provide quantitative information for various climate variables (e.g., sea surface temperature) over a given period. These datasets serve multiple purposes: they are used to develop and assess physical understanding of the climate system, to establish a baseline for evaluating changes in the climate system, and to develop, refine, and calibrate climate models.

Observational data are data that can be observed and measured. Much of the observational data used in climate research comes from instrumental records of climate variables. Examples include ground measurements of CO₂ concentrations, surface temperatures, and sea levels; satellite measurements of sea surface temperature, water vapor, precipitation, and sea ice; and aircraft measurements of cyclone wind speed.⁴¹ The first quasi-global instrumental datasets for land and sea surface temperature can be traced back to the mid-1800s (when national

40. See, e.g., Feng Zhu et al., *Climate Models Can Correctly Simulate the Continuum of Global-Average Temperature Variability*, 116 PNAS 8728 (2019), <https://doi.org/10.1073/pnas.180995911>.

41. Instrumental records are sometimes described as “direct” measurements of climate variables, but the directness of the measurement depends on the instrument being used. For example, surface thermometers provide relatively direct measurements of temperatures, whereas satellite microwave sensors record microwave emissions from which scientists can derive measurements of climate variables such as temperature and columnar water vapor. It is also important to keep in mind that instrumental records are not perfect—they may be subject to errors (e.g., calibration errors, sources of systematic error in the instrument, or interpretation techniques). However, scientists use calibration, validation, and verification techniques, often involving cross-comparison between different datasets, to ensure the accuracy of instrumental records.

and colonial weather services built networks of surface stations and began maintaining standardized weather logs).⁴² Since then, the instrumental record has grown significantly as a result of the expansion of ground measurement stations, as well as the advent of satellite remote sensing of climate variables in the 1980s. The instrumental record of current climate variables is now fairly comprehensive, but there are still spatial and temporal gaps in the instrumental record, particularly with regard to historical climate conditions and sea-level rise.

Paleoclimate reconstructions are used to evaluate climate conditions in periods prior to instrumental records. Paleoclimate research uses geological and biological evidence to reconstruct historical climate conditions for which direct measurements are not available. Paleoclimate researchers can obtain data on past climate conditions (e.g., precipitation and temperature) from sediments, rocks, tree rings, corals, ice sheets, and other physical formations. Paleoclimate reconstructions provide insights on baseline climate conditions and natural variability in those conditions before humans began influencing the atmosphere or using instruments to measure different climate variables.

Statistical Techniques and Climate Models

Statistical analysis refers to the mathematical formulas and techniques that are used in empirical analysis of data. Scientists use statistical techniques to fill gaps in observational datasets. Reanalysis datasets are generated by combining available observations with statistical analyses and climate models to infer climate conditions where direct observational data are not available. For example, interpolation can be used to fill in gaps in space and time for a climate variable based on the locations and times for the data that are available for the given variable, and thus plays a key role in data reanalysis.⁴³ Reanalysis datasets can be used to create “a coherent, long-term record of past weather by compensating for the inherent limitations of the different instruments used to take measurements at different points in the history of weather observation.”⁴⁴

Statistical analysis is also used in conjunction with observational data and climate models to detect changes in the climate system and other interconnected systems, is used to determine whether these changes are attributable to human influence (“detection and attribution” research),⁴⁵ and is used to predict

42. IPCC AR6 WG1, *supra* note 6, at 175.

43. Statistical interpolation involves estimating an unknown value based on related known values. There are many interpolation techniques, but the simplest example is linear interpolation—i.e., if observations show a linear increase in a variable over time, scientists can infer the value of that variable at a specific point in time even without a direct measurement.

44. National Oceanic and Atmospheric Administration National Centers for Environmental Information, *Reanalysis*, <https://www.ncei.noaa.gov/products/weather-climate-models/reanalysis>.

45. See section titled “Detection and Attribution” below.

future changes based on different warming trajectories.⁴⁶ For example, scientists will use statistical trend detection to determine whether there have been statistically significant changes in climate variables (e.g., precipitation) and related variables (e.g., crop yield).⁴⁷

Climate models use quantitative methods, including predictive equations and statistical techniques, to simulate interactions within the climate system. Scientists can set up different model experiments to evaluate the effect of one or more climate drivers (e.g., GHGs, aerosols, and solar flux) on one or more variables. As with statistical analysis, climate models can be used for the purposes of climate change detection and attribution as well as the projections of future climate change. Generally speaking, climate models are based on “well-established physical principles and have been demonstrated to reproduce observed features of recent climate . . . and past climate changes.”⁴⁸ Well-understood physical laws that are solved mathematically through space and time in climate models include the conservation of mass and energy. Indeed, many aspects of climate change can be modeled with a fair degree of accuracy, particularly at continental and global scales.⁴⁹ Climate models reproduce key planetary features like jet streams and ocean currents, as well as the relatively rapid warming of the northern hemisphere high latitudes in recent decades. Furthermore, a recent study found that past climate models had accurately predicted subsequent increases in global mean surface temperature (GMST), particularly when accounting for differences in projected and actual future forcings.⁵⁰ However, because of the complexity of the climate system with its nonlinear interactions and hyperlocal nature of important physical processes (e.g., thunderstorms, or shear stress at the edge of an ice sheet), there are limitations in the ability of models to accurately and precisely simulate causal relationships and processes, particularly when predicting effects of climate change at smaller spatial and/or temporal scales. Many of these complexities are partially addressed using parameterizations, which are simplified statistical estimates of complex phenomena.

There are several types of climate models with different applications, including the following: global circulation models (GCMs), which simulate general circulation of the Earth’s atmosphere and oceans; energy balance models (EBMs), which simulate changes in temperature based on the balance between incoming

46. See section titled “Projections of Future Climate Change” below.

47. R.K. Jaiswal et al., *Statistical Analysis for Change Detection and Trend Assessment in Climatological Parameters*, 2 *Env’t Processes* 729 (2015), <https://doi.org/10.1007/s40710-015-0105-3>.

48. David A. Randall et al., *Climate Models and Their Evaluation*, in IPCC, *Climate Change 2007: The Physical Science Basis*, Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change 591 (Susan Solomon et al. eds.), <https://perma.cc/GG2X-AEPE>.

49. See, generally, *id.* at 589–648.

50. Zeke Hausfather et al., *Evaluating the Performance of Past Climate Model Projections*, 47 *Geophysical Resch. Letters* e2019GL095378 (2020), <https://doi.org/10.1029/2019GL085378>.

and outgoing solar radiation; earth system models (ESMs), which simulate the interactions of the atmosphere, hydrosphere, geosphere, biosphere, and cryosphere to assess how changes in climate (e.g., warmer temperatures) affect and are affected by other interconnected systems such as changes in vegetation; and regional climate models (RCMs), which simulate the interactions between the climate and other earth systems at a finer resolution than ESMs. Coupled GCM and ESM models use understanding of thermodynamics, fluid motion, and other physical processes to simulate interactions between multiple components of the climate system, including atmospheric and ocean circulation. The Coupled Model Intercomparison Project (CMIP) of the World Climate Research Programme is an international initiative that combines, synthesizes, and compares different climate models in order to develop more accurate climate simulations. The CMIP model ensembles are updated periodically—the latest IPCC assessment uses results from CMIP Phase 6 (CMIP6) models, which “include new and better representations of physical, chemical, and biological processes, as well as higher resolution” compared to past climate models.⁵¹

Generally speaking, model-based approaches can support more robust findings than the use of observational data and statistical analysis alone. However, models have limitations that should be kept in mind when evaluating their results. The usefulness and accuracy of a model depends on how well the model can reproduce patterns associated with each climate forcing, and there are uncertainties in our knowledge about how individual climate forcings affect different aspects of the climate system. Comparing model results to observations can help assess the accuracy of the model, but observations cannot tell us all we need to know for several reasons. First, there is uncertainty in observational measurements for reasons discussed above. Second, internal climate variability, unrelated to climate forcing, is difficult to disentangle from climate forcing. Third, because multiple anthropogenic and natural forcings have occurred simultaneously in the past, unpacking the relative contribution of each forcing is a major challenge.

The above challenges exist to a certain degree even for variables like global average temperature where the relationship between rising GHG concentrations and average temperatures is fairly direct. Inevitably, there will be some degree of uncertainty and room for error in model results due to the complexity of the physical systems being modeled, so scientists have tools for managing and communicating uncertainty and error rates.⁵² Scientists are also constantly refining the techniques used to reduce uncertainty in their analyses, such as

51. According to the IPCC, the use of CMIP6 “has improved the simulation of the recent mean state of most large-scale indicators of climate change and many other aspects across the climate system.” However, there are still some differences between CMIP6 model results and observations, particularly with regard to regional precipitation. IPCC AR6 WGI, *supra* note 6, at 12.

52. See section titled “Managing and Communicating Uncertainty” below.

through additional and lengthened observational datasets, improvements to models, new analytical methods, and expert judgment. For example, new statistical approaches are being used to better account for internal climate variability and uncertainties in models and observations.⁵³

Another limitation to GCMs and other large-scale models is that they produce results at large spatial scales and thus cannot simulate local climate features and their impacts with precision. To address this issue, researchers are developing dynamical and statistical downscaling techniques that can be used to transform climate model outputs into more localized data products for regional or local climate impact assessments, although these approaches can run up against inherent uncertainties at fine spatial scales.⁵⁴

Managing and Communicating Uncertainty

Owing to the complexity of the climate system, researchers inevitably confront uncertainty when evaluating causal relationships and processes within the system. This is not unique to the field of climate science: uncertainty exists across all scientific disciplines, and understanding sources of uncertainty is part of the scientific process. Climate scientists use standard techniques and practices for managing and communicating uncertainty and ensuring the validity of research findings. These include statistical- and model-based approaches that actually reduce uncertainty in research findings,⁵⁵ as well as methods of framing and communicating uncertainty along with findings.

The IPCC, for example, uses probabilistic language to describe the assessed likelihood of an outcome or result and uses other language to communicate confidence and level of agreement in findings. Specifically, the IPCC describes (1) the assessed likelihood of an outcome or result (very likely, likely, etc.); (2) the availability of evidence to support particular findings (limited, medium, robust); (3) the level of agreement about findings (low, medium, high); and (4) scientific confidence in the findings (very low, low, medium, high, very high), which is based on both the level of agreement and availability of evidence for the

53. IPCC AR6 WGI, *supra* note 6, at 205. See also *id.* § 3.2.

54. “Dynamical downscaling refers to the use of high-resolution regional simulations to dynamically extrapolate the effects of large-scale climate processes to regional or local scales of interest. Statistical downscaling encompasses the use of various statistics-based techniques to determine relationships between large-scale climate patterns resolved by global climate models and observed local climate responses.” NOAA Geophysical Fluid Dynamics Lab’y, Climate Model Downscaling, <https://perma.cc/7ZBZ-NQ9D>.

55. See, e.g., Flavio Lehner et al., *The Potential to Reduce Uncertainty in Regional Runoff Projections from Climate Models*, 9 Nature Climate Change 926 (2019), <https://doi.org/10.1038/s41558-019-0639-x>.

finding.⁵⁶ The full list of terms used to communicate likelihood in the most recent IPCC report is as follows: virtually certain, 99–100% probability; very likely, 90–100%; likely, 66–100%; about as likely as not, 33–66%; unlikely, 0–33%; very unlikely, 0–10%; and exceptionally unlikely, 0–1%. Additional terms (extremely likely, 95–100%; more likely than not, >50–100%; and extremely unlikely, 0–5%) are also used when appropriate.⁵⁷ The use of such calibrated uncertainty language can make scientific findings more accessible to scientists and nonscientists alike.

Importantly, a finding of “low evidence” or “low confidence” does not equate to a finding that a particular proposition is not true or valid—it simply means that there is not enough evidence for IPCC scientists to reach agreement on the proposition. As new scientific data become available for subsequent assessments, the IPCC often revises these statements to reflect greater levels of confidence (e.g., the IPCC expressed greater confidence in the attribution of extreme events to climate change in AR6 than it had in past assessments).⁵⁸

In individual studies, uncertainty is typically managed using similar statements about probabilities as well as confidence levels and intervals. A confidence interval reflects a range of possible true values for the parameter being studied. Confidence intervals are typically expressed as the mean estimate of the parameter $\pm$ variation in the estimate at a designated confidence level (typically 95%, although 90% and 99% are also used). The confidence level reflects the likelihood that the parameter will fall within the upper and lower bounds of the confidence interval. For example, a study may conclude with 95% confidence that anthropogenic climate forcing increased the likelihood of a specific extreme event by a factor of 4 ± 1 (with $4x$ being the mean estimate, and $3\text{--}5x$ being the confidence interval). Some studies will present findings at lower confidence bounds to provide additional insights on likely or probable findings (e.g., a 66% confidence level, corresponding with a “likely” finding).⁵⁹

Scientific studies also typically include information about Type I (false positive) and Type II (false negative) errors. If Type I errors are high, then the study may have produced a spurious association between anthropogenic forcing and an observed trend. Conversely, if Type II errors are high, then the study may be underestimating or wholly missing the effects of anthropogenic forcing on an observed trend. Climate researchers, and academic researchers more generally, tend to be more concerned with avoiding Type I errors to ensure that they do

56. A. Kause et al., *Confidence Levels and Likelihood Terms in IPCC Reports: A Survey of Experts from Different Scientific Disciplines*, 173 *Climatic Change* 1 (2022), <https://doi.org/10.1007/s10584-022-03382-3>.

57. IPCC AR6 WG1, *supra* note 6, at 4.

58. See section titled “Extreme Event Detection and Attribution” below.

59. See, e.g., Mark D. Risser & Michael F. Wehner, *Attributable Human-Induced Changes in the Likelihood and Magnitude of the Observed Extreme Precipitation During Hurricane Harvey*, 44 *Geophysical Resch. Letters* 12,457 (2017), <https://doi.org/10.1002/2017GL075888>.

not overstate the magnitude of anthropogenic climate change. Indeed, the high burden of proof assumed in standard statistical tests leads to researchers being very conservative in their estimates of climate change and its effects.⁶⁰

A metric known as the p -value provides further insights on the validity of climate studies. The p -value is the quantification of the probability of a Type I error, such as the probability that an observed trend such as global surface warming would occur due to chance alone. A p -value of 5% or less is commonly used as a threshold of validity in climate studies and other areas of natural science.⁶¹ The frequent use of such a low p -value thus reflects the aversion of scientists to false positives/Type I errors.

Sources of Climate Research

Because of the breadth and complexity of climate science, scientific organizations like the IPCC play an important role in the synthesis and dissemination of climate research. The U.S. federal government also funds research and publishes reports on climate science, and there are thousands of individual researchers, academic institutions, and NGOs contributing to this field. This section provides context on different sources of climate research and highlights some considerations to help judges assess the credibility, weight, and admissibility of research depending on the source. The Appendix to this reference guide provides additional information about some of the organizations and sources discussed herein.

Scientific and Consensus Reports

There are several major organizations that periodically publish reports on the state of climate science. These include the IPCC, the World Meteorological Organization (WMO), the American Meteorological Society (AMS), the American Geophysical Union (AGU), the National Academies of Sciences, Engineering, and Medicine (National Academies or NASEM), and the U.S. Global Change Research Program (USGCRP). The WMO, for example, is a leading source of climate data products and it publishes an annual *State of the Global Climate Report*

60. William R.L. Anderegg et al., *Awareness of Both Type I and 2 Errors in Climate Science and Assessment*, 95 Bull. Am. Meteorological Soc'y 1445 (2014), <https://doi.org/10.1175/BAMS-D-13-00115.1>.

61. A p -value less than or equal to 5% is often described as the threshold for statistical significance, but this does not relate to the magnitude of impact—rather it is associated with probability. If the p -value is >5%, then there is a reasonable probability that an observed trend (for example) might be due to chance alone.

that summarizes the latest observations and findings for various global climate indicators including GHGs, global temperature, ocean heat content, sea level, marine heat waves, the cryosphere, and precipitation.⁶²

The IPCC is widely considered to be the leading scientific body for the assessment and synthesis of research on climate change. The IPCC does not conduct new research. Rather, it publishes assessment reports based on a synthesis of thousands of published, peer-reviewed studies from scientific journals.⁶³ These assessment reports are prepared with input from thousands of scientists with diverse expertise across the field of climate research.⁶⁴ A key benefit of the IPCC reports is that they identify scientific findings that have multiple lines of evidence, have been replicated, and have stood the test of time. There is a robust process for analyzing existing research and reaching conclusions on the basis of the reviewed science, as detailed in the Appendix to this reference guide. Thus, the IPCC reports and findings contained therein reflect a level of scientific scrutiny and agreement that is unique in this field.

Taking into account the procedures underpinning IPCC reports, the U.S. Supreme Court and federal appellate courts have recognized these reports as an authoritative and credible source of climate science.⁶⁵ However, it is possible that judges may confront disputes regarding the accuracy of IPCC findings, particularly if there is more recent and credible scientific research that calls those findings into question.⁶⁶ Although IPCC reports are typically afforded greater weight in the scientific community than individual studies, the science is constantly evolving, and subsequent research may provide new insights on the nature of climate change and its consequences. This progression toward greater scientific confidence

62. See WMO, State of the Global Climate, <https://perma.cc/B4X9-9SDV>.

63. The reports also draw on so-called “grey literature” (i.e., non-peer-reviewed reports, including technical reports, conference proceedings, statistics, and observational datasets), but there are strict guidelines for its inclusion.

64. The scientists involved in the assessments are selected through a nomination process. Prior to initiating an assessment, the IPCC issues a call to governments and IPCC observer organizations for nominations; the authors are then selected by the Bureau of Scientists on the basis of their expertise. The IPCC seeks to build author teams that reflect a range of scientific, technical, and socioeconomic expertise. See IPCC, Factsheet: How Does the IPCC Select Its Authors? (2021), <https://perma.cc/8MGL-R9LL>.

65. See *Massachusetts v. EPA*, 549 U.S. 497 (2007); *Coal. for Responsible Regul. v. EPA*, 684 F.3d 102 (D.C. Cir. 2012); *Ctr. for Biological Diversity v. Nat’l Highway Traffic Safety Admin.*, 538 F.3d 1172, 1189 (9th Cir. 2008); *Diné Citizens Against Ruining Our Env’t v. Haa-land*, 59 F.4th 1016 (10th Cir. 2023).

66. By “more recent” research, we mean research that was published after information was collected for the latest IPCC assessment (and therefore could not have been incorporated into the assessment).

in both the attribution and prediction of climate change impacts is evident across IPCC assessments.⁶⁷

There are other synthesis reports that serve as important sources of climate data; some of these reports provide more targeted assessments of specific topics and/or geographic regions. For example, the USGCRP periodically publishes *National Climate Assessments* (NCAs) that integrate, evaluate, and interpret scientific findings related to climate change and its effects on the United States, including effects on the natural environment, agriculture, energy production and use, land and water resources, transportation, human health and welfare, human social systems, and biological diversity. The reports also analyze broader trends in global climate change, both human-induced and natural, and projections for the subsequent 25 to 100 years.

Peer-Reviewed Research

The assessments performed by the IPCC and other authoritative science bodies are based primarily on syntheses of peer-reviewed climate research. Litigants may also rely on individual peer-reviewed studies and articles to support scientific claims. The benefit of peer review is that it ensures that research has been examined by one or more scholars with expertise in the subject matter (although it does not generally involve repeating any of the measures or calculations or otherwise reproducing the work). Examples of peer-reviewed research include original research studies (i.e., primary research), review articles, and expert judgment reports.⁶⁸ The most robust climate studies tend to be those that combine good observational data, physical understanding, rigorous statistical analysis, and detailed models to generate findings, along with clear communication and transparency with respect to research parameters, assumptions made, confidence in findings, and potential areas of uncertainty or bias.

Importantly, different publications have different standards for what qualifies as peer review, and there are some journals that purport to publish peer-reviewed research that do not actually have a legitimate peer-review process—these are sometimes referred to as “predatory journals.” In order to distinguish between legitimate and illegitimate publications, judges can refer

67. See, e.g., IPCC AR6 WGI, *supra* note 6, at 52 (“new techniques developed since AR5, including attribution of individual events, have provided greater confidence in attributing changes in climate extremes to climate change”).

68. Here, “expert judgment reports” specifically refers to peer-reviewed articles containing findings based on expert judgment and expert surveys (not to be confused with expert reports submitted as part of litigation). See, e.g., Jonathan L. Bamber et al., *Ice Sheet Contributions to Future Sea-Level Rise from Structured Expert Judgment*, 116 PNAS 11195 (2019), <https://doi.org/10.1073/pnas.1817205116>.

to online lists of such journals;⁶⁹ judges can also examine the credentials of the journal editors, the authors of the particular study at issue in the case, and the authors of other studies published in that same journal.⁷⁰ Another indication of the quality of a journal is whether the publication has been indexed by major journal-indexing groups such as Scopus, Web of Science, PubMed, and Google Scholar.

Individual studies and reports are typically not afforded the same weight within the scientific community as IPCC assessments and other major scientific reports, but they can serve as important supplements to such reports, as they may provide insights on areas of climate science that are not covered elsewhere (e.g., assessments of climate change impacts at a more local or granular scale, or findings based on data that are too recent to have been included in a prior synthesis report).

It is also important to note that the fact that a scientific resource is *not* peer reviewed does not in and of itself mean that the resource is faulty or illegitimate. There are credible scientific data and research products that do not undergo formal peer review—this includes, for example, some of the data and research products published by government agencies. Courts can consider other factors when evaluating the credibility of such resources, including the credentials of the publishing organizations and scientists, whether the findings are consistent with those from expert bodies like the IPCC, and whether the underlying methodologies have been subject to peer review.

Expert Testimony and Reports

Expert witnesses can play an important role in communicating and interpreting scientific evidence in climate litigation. Part of this role may involve simply summarizing findings from the IPCC and other authoritative bodies and explaining the relevance of those findings to a particular case. Expert witnesses may be needed to support or refute factual claims that are beyond the scope of broad-scale climate assessments like IPCC reports—e.g., claims about the effects of climate change on a particular locale or individual. When that is the case, expert witnesses may provide testimony and reports based on individual studies or impact assessments (including government assessments). They can also answer questions to help clarify the methods and findings from specific studies—for example, explaining why a particular time frame or historical baseline was used

69. See, e.g., Beall's List of Potentially Predatory Journals and Publishers, <https://perma.cc/63GD-ZS25>.

70. See section titled “Expert Testimony and Reports” for further insights on engaging with expert witnesses.

in a study, or explaining the level of uncertainty inherent in a particular finding.

There are contexts where expert witnesses may draw inferences about the effects of climate change in the absence of a targeted study on those effects.⁷¹ For example, an expert may infer that climate change has contributed to more severe heat waves in a particular location based on regional analyses of climate impacts and physical understanding of the strong causal relationship between climate change and extreme heat.⁷² The reasonableness of such inferences would depend on factors such as the nature and location of the impact, the strength of the “signal” of anthropogenic climate change relative to natural variability, and the level of spatial or temporal variability in the impact. In particular, when evaluating the potential causal link between climate change and a specific event or impact, a judge could consider whether (1) a widespread pattern in a geographic area has been observed and attributed to climate change, (2) there is not too much spatial or temporal variability in the impact, and (3) the expert is inferring that this pattern applies to a particular locale.

There are some threshold considerations when assessing the admissibility of expert testimony on climate science.⁷³ First, the field of climate science is so broad that it is impossible to articulate general criteria for expert qualifications in this field—whether a witness is qualified to speak will be a case-by-case determination that is entirely dependent on the scope of their testimony. For example, an expert who is testifying on the reliability of global climate models should have expertise on the physical processes that drive global climate change (which could be gleaned through, e.g., research in atmospheric sciences or meteorology) and/or the statistical and mathematical techniques deployed in those climate models; an expert who is testifying on the biological impacts of climate change should have expertise in biological sciences; and so forth.

Second, when evaluating whether the testimony would pass the *Daubert* test (i.e., whether it is based on reliable principles and methods), courts may consider factors such as: (1) whether the testimony is based on principles,

71. See Elisabeth A. Lloyd & Theodore G. Shepherd, *Climate Change Attribution and Legal Contexts: Evidence and the Role of Storylines*, 167 *Climatic Change* 27 (2021), <https://doi.org/10.1007/s10584-021-03177-y>. The authors note that “proceeding from the general to the specific is a process of deduction and is an entirely legitimate form of scientific reasoning” and “well aligned with the concept of legal evidence.” *Id.* (abstract).

72. See section titled “Extreme Event Detection and Attribution” below.

73. As discussed in Liesa L. Richter and Daniel J. Capra, *The Admissibility of Expert Testimony*, in this manual, when such disputes arise, judges must find that the expert witness is qualified to provide testimony on the subject matter, that the witness’s scientific and technical knowledge will help the trier of fact understand the subject matter, that the expert’s opinion is based on sufficient facts or data, that the expert’s opinion is the product of reliable principles and methods, and that the witness has reliably applied those principles and methods to the facts of the case. See also Fed. R. Evid. 702; Fed. R. Evid. 104(a).

methodologies, or findings that have been accepted as credible by the IPCC and other scientific institutions; (2) whether the underlying methods and findings have been subjected to peer-review processes; and (3) whether the research is accompanied by information about confidence levels and error rates.⁷⁴ When confronted with novel research findings or methodologies, judges can consider whether the novel aspects are rooted in existing and accepted scientific techniques to determine whether they represent a significant departure from general practices. In many cases, so-called “novel” techniques used in climate studies are based on minor changes to (or advances in) well-established research methods (and in some cases, they are adapted from other disciplines). For example, extreme-event attribution generally relies on the same climate models used in past studies to attribute changes in average conditions and to predict future changes,⁷⁵ and probabilistic or risk-based extreme-event studies also use concepts and methods developed in epidemiological research.⁷⁶

Climate Change Detection, Attribution, and Projections

Litigants typically use climate science to support or refute claims about the causes and impacts of climate change. For example, a plaintiff may use climate science to demonstrate that the GHG emissions attributable to a defendant’s conduct have caused or contributed to an injury incurred by the plaintiff or society at large. There are several areas of climate research that are particularly relevant to litigation. First, detection and attribution research examines whether and to what extent specific trends, events, and impacts can be attributed to human influence on the climate system. Second, source attribution research evaluates the respective contributions of different actors, activities, sectors, and jurisdictions to anthropogenic climate change and its impacts. Third, projections of future climate change

74. As discussed above, studies often include information about Type I (false positive) and Type II (false negative) errors, as well as the *p*-value (the probability of obtaining results at least as extreme as the observed result, assuming that the null hypothesis is correct). A 5% *p*-value cutoff is commonly used to ensure the validity of results. In addition, climate studies and IPCC reports use confidence levels to communicate the likelihood that a finding is valid, and most individual studies use a high confidence level (e.g., ≥ 90%) corresponding with a low margin of error. These confidence statements may not be directly relevant to the *Daubert* inquiry insofar as they deal with the validity of findings rather than methodologies. However, they are relevant when assessing the credibility and probative value of expert testimony. See *Daubert v. Merrell Dow Pharms., Inc.*, 509 U.S. 579 (1993) (listing nonexhaustive factors in determining reliability of a scientific expert method).

75. See section titled “Impact Detection and Attribution” below.

76. See Theodore G. Shepherd, *A Common Framework for Approaches to Extreme Event Attribution*, 2 Current Climate Change Reps. 28 (2016), <https://doi.org/10.1007/s40641-016-0033-y>.

provide insights on the scope and magnitude of future climate change under different warming and emissions trajectories. Finally, there is research aimed at estimating remaining carbon budgets that would limit global warming to targets such as 1.5°C, 2.0°C, or “well below” 2.0°C. For each of these research areas, we describe the underpinning methods and parameters, and we summarize key research findings, focusing on findings from the IPCC and other scientific bodies.

As detailed below, there is scientific consensus on the reality of anthropogenic climate change, and scientists can detect, attribute, and predict many of the trends and impacts caused by climate change with a high level of confidence. But some gaps remain in scientific knowledge of the climate system and uncertainty about the effects of climate change, particularly when looking at the regional or local impacts of climate change.

Detection and Attribution

Detection and attribution research examines the causal links between human activities, changes in the climate system, and corresponding impacts on other interconnected systems.⁷⁷ Research in this field has demonstrated that human activities are the dominant cause of observed climate change,⁷⁸ and that human-induced climate change is causing pervasive impacts to human and natural environments around the world.⁷⁹ Some of the observed changes include rising sea levels, ocean warming and acidification, melting sea ice, thawing permafrost, increases in the frequency and severity of many types of extreme events, and corresponding impacts on people, communities, and ecosystems.⁸⁰ These findings are based on multiple lines of evidence, including physical understanding of the climate system and the greenhouse effect, comparisons between observational data and climate models, paleoclimate reconstructions, and “fingerprinting”

77. This discussion of attribution research has been adapted and, in some cases, excerpted from the authors’ prior publication on this topic. See Michael Burger et al., *The Law & Science of Climate Change Attribution*, 5 Colum. J. of Env’t Law 57 (2020), <https://perma.cc/N7A7-4XMP>. The discussion of scientific findings has been updated to reflect new resources, including the latest IPCC assessment (AR6).

78. NAS Update, *supra* note 6, at ch. 2. See also IPCC AR6 WGI, *supra* note 6; NCA4 Vol. I, *supra* note 6.

79. The weight of the scientific evidence demonstrates that anthropogenic climate change “is already affecting every inhabited region across the globe, with human influence contributing to many observed changes in weather and climate extremes.” IPCC AR6 WGI, *supra* note 6, at 10. See also NCA4 Vol. I, *supra* note 6, at 36 (“Evidence for a changing climate abounds, from the top of the atmosphere to the depths of the oceans.”).

80. See sections titled “Source Attribution” and “Projections of Future Climate Change” below.

studies that examine the influence of anthropogenic forcing on specific climatological trends and events.⁸¹

Detection and attribution research can be categorized into different subfields, each of which corresponds with a different link in the causal chain connecting human activities to climate-change-related harms:

- **Detection and attribution of climate change** focuses on the link between anthropogenic climate forcing and corresponding changes in the climate system, including the atmosphere, hydrosphere, cryosphere, land surface, and biosphere.
- **Extreme-event attribution** examines how global climate change has affected the probability, frequency, severity, and other characteristics of extreme events such as heat waves, storms, floods, droughts, and wildfires.⁸²
- **Impact attribution** examines how global climate change is affecting human and natural systems. This research deals with a broad range of physical, social, health, economic, and biological impacts at global, regional, and local scales.⁸³

Below, we provide a general overview of methods and parameters used in attribution research. This is followed by a more targeted discussion of the three subfields identified above (climate change, extreme-event, and impact attribution).

81. IPCC AR6 WGI, *supra* note 6; NAS Update, *supra* note 6, at ch. 2. For additional information on fingerprinting studies, see generally this section and the subsection titled “Attribution to anthropogenic climate change” below.

82. We discuss extreme-event attribution as a separate category of attribution research because extreme events do not fit neatly into the “global climate change” or “impact” attribution categories. Weather is part of the climate system, but extreme events are often discussed as impacts of climate change, and there are unique challenges associated with efforts to ascertain the effect of climate change on a particular extreme event.

83. The distinction between “changes in the global climate system” and “the impacts of climate change” is not always clear because of the broad definition of the global climate system. The IPCC defines *impacts* or *effects* to include physical impacts such as floods, droughts, and local sea-level rise, as well as any other “effects on lives, livelihoods, health and well-being, ecosystems and species, economic, social and cultural assets, services (including ecosystem services), and infrastructure.” IPCC AR6 WGI, *supra* note 6, at 201. In many cases, a change in an essential climate variable (e.g., sea-level rise) could be viewed as a physical impact of climate change. For the purposes of this reference guide, we classify studies on regional changes in essential climate variables as “climate change attribution” where the primary goal of the study is to better understand how humans are affecting the global climate system, and we classify studies on floods, droughts, and local sea-level rise as “impact attribution,” where the primary goal of the study is to better understand how climate change is affecting a particular region or locale.

General Methods and Parameters

Detection of change

Detection and attribution is a two-step process used to identify a causal relationship between one or more drivers and a responding system. The first step—detection of change—involves demonstrating that a particular variable has changed in a statistically significant way without assigning cause.⁸⁴ To accomplish this, scientists will compare historical climate data with contemporary observations to assess the magnitude of change, and they will also evaluate whether the observed change may be due to internal variability or external forcing on the climate system. An identified change is detected in observations if the likelihood that it occurred because of internal variability (i.e., chance) alone is determined to be small, for example, less than 10%.⁸⁵

Scientists typically use instrumental records to identify an event or process that will be the subject of the detection and attribution study. The subject could be a gradual process, such as increases in average sea surface temperature or global mean sea level, or it could be a sudden-onset event, such as a heat wave. Scientists will also use instrumental records in conjunction with other sources of climate data (e.g., reanalysis datasets and paleoclimate reconstructions) to characterize baseline climate conditions and to evaluate how those conditions have changed over time.

Attribution to anthropogenic climate change

The second step—attribution—involves sifting through a range of possible causative factors to determine the role of one or more drivers with respect to the detected change. This is typically accomplished by using physical understanding, as well as climate models and/or statistical analysis, to compare how the variable responds when certain drivers are changed or eliminated entirely. The goal of such studies is to determine whether, how, and to what extent anthropogenic drivers have contributed to the observed change.

Many attribution studies use a probabilistic approach—i.e., researchers will seek to quantify the probability of a particular outcome (e.g., how likely is the occurrence of three inches of rainfall in a day at a given locale) occurring with and without anthropogenic influence on climate. However, researchers can

84. David R. Easterling et al., *Detection and Attribution of Climate Extremes in the Observed Record*, 11 *Weather & Climate Extremes* 17 (2016), <https://doi.org/10.1016/j.wace.2016.01.001>; Gabriele Hegerl, *Towards Detection and Attribution of Impact-Relevant Climate Change: The WG1 Perspective*, in *IPCC Expert Meeting on Detection and Attribution Related to Anthropogenic Climate Change*, Meeting Report 25–27 (Thomas Stocker et al. eds., 2010), <https://perma.cc/X3HY-VR9L>.

85. IPCC AR6 WGI, *supra* note 6, at 196.

also use a mechanistic approach to attribution, whereby they seek to examine how climate change has influenced one or more physical characteristics of an event or process.⁸⁶ Mechanistic studies can provide insights on, for example, the change in magnitude or severity of an extreme event that can be attributed to climate change.⁸⁷ In the rainfall example above, a mechanistic approach might look at the weather system that produced the heavy rain and describe how one part of climate change that we understand well—e.g., the warming of the atmosphere and its resulting increase in the amount of moisture the atmosphere can hold—contributed to the event. Mechanistic and probabilistic analyses can be combined in order to develop a more complete picture of whether and to what extent climate change is influencing various processes and events.⁸⁸

Researchers use both statistical techniques and climate models when detecting and attributing change. As an example of statistical techniques, scientists use linear regression methods⁸⁹ and variants such as “optimal fingerprinting” to determine whether a change in a climate variable is statistically significant or simply part of natural variability.⁹⁰ This analysis is part of the detection of climate change and corresponding impacts, but it can also be used to support attribution statements (e.g., a finding that the spatial pattern of warming in the atmosphere was likely caused by anthropogenic emissions because it is statistically unlikely that the spatial pattern would have occurred in the absence of anthropogenic forcing on the climate). This is sometimes referred to as observation-based attribution.⁹¹

86. See section titled “Projections of Future Climate Change” below for further discussion of the mechanistic approach and its role in extreme-event attribution (also referred to as the “story-line” approach to extreme-event attribution).

87. See, e.g., Michael Wehner & Christopher Samson, *Attributable Human-Induced Changes in the Magnitude of Flooding in the Houston, Texas Region During Hurricane Harvey*, 166 *Climatic Change* 19 (2021), <https://doi.org/10.1007/s10584-021-03114-z>. See also Luke J. Harrington et al., *Integrating Attribution with Adaptation for Unprecedented Future Heatwaves*, 172 *Climatic Change* 1 (2022), <https://doi.org/10.1007/s10584-022-03357-4> (discussing the differences and similarities between probabilistic and mechanistic approaches to extreme-event attribution).

88. For example, findings from both probabilistic and mechanistic studies are synthesized in IPCC assessments and other climate reports.

89. Linear regression is a statistical method used to summarize and study relationships between two continuous (quantitative) variables.

90. Optimal fingerprinting regresses observed climate variables on expected responses to, or signals of, specific forcings to determine whether and to what extent the signals are present in the observation. See Zhuo Wang et al., *Toward Optimal Fingerprinting in Detection and Attribution of Changes in Climate Extremes*, 116 *J. Am. Stat. Ass’n* 1 (2021), <https://doi.org/10.1080/01621459.2020.1730852>; K. Hasselmann, *Optimal Fingerprints for the Detection of Time-Dependent Climate Change*, 6 *J. Climate* 1957 (1993), [https://doi.org/10.1175/1520-0442\(1993\)006<1957:OFFTDO>2.0.CO;2](https://doi.org/10.1175/1520-0442(1993)006<1957:OFFTDO>2.0.CO;2).

91. Nat’l Acad. of Sci., Eng’g, & Med. (NASEM), *Attribution of Extreme Weather Events in the Context of Climate Change* (2016), <https://perma.cc/D56R-G9DJ>.

However, in practice, most studies do not rely exclusively on observation-based statistical analysis for attribution because of short observation records and complex forcing changes over the historical period.⁹² Climate models are typically used for attribution because they allow scientists to separate out the effects of different forcings and processes on the observed variable. That said, observation-based attribution findings can serve as a useful supplement to model-based findings.⁹³

Attribution studies utilizing climate models generally involve at least two sets of simulations: one that reflects the actual world, and another that reflects a counterfactual world without anthropogenic climate change (or without some component of anthropogenic climate change). These model simulations are ideally run at least several times based on differing initial conditions and for long duration, allowing scientists to better differentiate the climate change signal from the noise of natural variability. Observational data and physical understanding provide the basis for calibrating and verifying models.

Several modeling centers have now developed standardized climate simulations designed for detection and attribution specifically, based on different parameters (e.g., researchers can evaluate the probability of an event or impact occurring both with and without certain observed changes in the climate, such as changes in sea surface temperature). Owing to advances in parallel computing and model simplifications, these can be run rapidly and at high spatial resolution, yielding quick results. Indeed, when the above packages are combined with forecasts of variables with high predictability, such as sea surface temperature, projected results can be made available *in advance* of actual events. Furthermore, the tools and outputs, and models themselves, are increasingly being made publicly available. This has furthered the proliferation of attribution research in recent years.

Challenges associated with downscaling and exogenous variables

Attribution becomes increasingly complex and challenging as the focus of research moves away from long-term, broad-scale changes in the climate system and toward more localized, discrete events and impacts. One key challenge is conducting the analysis at the appropriate spatial and temporal scale. Natural variability, unrelated to changes in anthropogenic climate forcing, is larger at fine spatial and temporal scales, making it harder to identify signals associated

92. *Id.*

93. Andrew D. King et al., *Attribution of the Record High Central England Temperature of 2014 to Anthropogenic Influences*, 10 *Env't Rsch. Letters* 1 (2015), <https://doi.org/10.1088/1748-9326/10/5/054002>; Gabriele C. Hegerl, *Use of Models and Observations in Event Attribution*, 10 *Env't Rsch. Letters* 1 (2015), <https://doi.org/10.1088/1748-9326/10/7/071001>.

with anthropogenic or other forcings.⁹⁴ When models are used to assess extreme events or impacts that occur at finer spatial and temporal scales than the models themselves, some type of downscaling or error correction is needed, which can introduce additional uncertainties.⁹⁵

Impact attribution studies must also account for nonclimate or exogenous variables, that is, characteristics of human and natural systems that are not part of the climate system.⁹⁶ Consider a study examining the relationship between climate change, a heat wave, and public health impacts: the study would need to account for both climate variables at a fairly discrete geographic and temporal scale (e.g., temperature and humidity during the event) as well as nonclimate variables (e.g., population risk factors for heat-related morbidity, access to air-conditioned facilities and emergency services) to ascertain the extent to which climate change caused or contributed to observed health outcomes. Confounding variables (e.g., air quality), which influence both dependent and independent variables in a study, are of special concern, as they can lead to spurious associations between a driver and an event or impact.⁹⁷ The number of exogenous and confounding variables increases as attribution research moves toward an analysis of discrete impacts on humans, communities, and ecosystems.

To manage exogenous variables, some studies use a single-step attribution approach, i.e., “a single modelling setup to relate changes in drivers to changes in some aspect of a climate, natural, or human system.”⁹⁸ As an example, single-step attribution has been used to evaluate the relationship between very warm regional temperatures and human influence on the climate.⁹⁹ Other studies use a multistep approach that links separate single-step analyses into an overall attribution assessment. Multistep studies often examine how one or more core climate variables have changed in response to human activities, and then explore

94. See IPCC AR6 WGI, *supra* note 6, at 117 (discussing how internal variability is stronger and uncertainties in observations, models, and external forcings are larger at the regional scale, as compared with the global scale).

95. See, e.g., NOAA Geophysical Fluid Dynamics Lab’y, Climate Model Downscaling, <https://perma.cc/7ZBZ-NQ9D>.

96. This may be somewhat of an oversimplification, as many variables that may appear to be outside of the climate system are still, to some extent, interdependent with that system.

97. In an impact or event attribution study, the dependent variable would be the impact or event under examination (e.g., a heat wave or an uptick in hospitalizations) and the independent variable would be the climate-change-related driver of the impact or event (e.g., increases in GHG concentrations or, in some studies, increases in climate variables such as mean temperature).

98. Dáithí Stone et al., *The Challenge to Detect and Attribute Effects of Climate Change on Human and Natural Systems*, 121 Climatic Change 381, 390–91 (2013), <https://doi.org/10.1007/s10584-013-0873-6>.

99. See Peter A. Stott et al., *Single-Step Attribution of Increasing Frequencies of Very Warm Regional Temperatures to Human Influence*, 12 Atmospheric Sci. Letters 220 (2011), <https://doi.org/10.1002/asl.315>.

the implications of that change with respect to one or more specific impacts.¹⁰⁰ Multistep attribution is useful for examining causal relationships in complex systems, but one potential drawback of this approach is that additional, cascading uncertainty and potential for error is introduced with each new step that is added to the analysis. Attribution researchers can also use an indirect two-step approach by referencing findings from other studies or IPCC reports to establish the first step in their causation analysis (e.g., the link between anthropogenic forcing and warmer average temperatures), and then they can focus on the second step (e.g., the link between warmer average temperatures and heat waves).¹⁰¹

Climate Change Detection and Attribution

Studies in this category examine the effects of anthropogenic climate forcing on specific components of the climate system and metrics such as average temperature and precipitation. This research provides foundational knowledge about global climate change that is subsequently used in extreme-event and impact attribution studies.

Methods and parameters

Scientists detect changes in the climate system through instrumental records, such as the CO₂ readings from the Mauna Loa Observatory in Hawaii, remote sensing from satellites, and other platforms. Some of the key variables being monitored include atmospheric concentrations of GHGs and other radiative forcers, atmospheric and surface temperature, water vapor (humidity), precipitation, sea ice, sea levels, ocean heat content, and ocean acidity. Scientists also use climate reanalysis datasets to fill gaps in instrumental records, as well as paleoclimate reconstructions to provide a perspective on longer-term climate change and variability. Paleoclimate reconstructions provide a means of comparing the current climate with that of past periods in Earth's history. The reconstructions

100. See, e.g., Yixin Mao et al., *Is Climate Change Implicated in the 2013–2014 California Drought? A Hydrologic Perspective*, 42 *Geophysical Research Letters* 2805 (2015), <https://doi.org/10.1002/2015GL063456>; A. Park Williams et al., *Contribution of Anthropogenic Warming to California Drought During 2012–2014*, 42 *Geophysical Research Letters* 6819 (2015), <https://doi.org/10.1002/2015GL064924>.

101. See, e.g., Peter Stott et al., *Future Challenges in Event Attribution Methodologies*, in *Explaining Extreme Events of 2016 from a Climate Perspective*, 99 *Bull. Am. Meteorological Soc'y* S1, S155 (2018), <https://doi.org/10.1175/BAMS-D-17-0118.1> [hereinafter BAMS 2016], referencing Russell E. Brainard et al., *Ecological Impacts of the 2015/16 El Niño in the Central Equatorial Pacific*, in BAMS 2016 at S21, <https://doi.org/10.1175/BAMS-D-17-0128.1>. See also David J. Frame et al., *The Economic Costs of Hurricane Harvey Attributable to Climate Change*, 160 *Climatic Change* 271 (2020), <https://doi.org/10.1007/s10584-020-02692-8>.

offer important insights, including: (1) how sensitive different aspects of the climate system are to different climate forcings at various timescales and (2) more robust estimates of natural variability than can be gleaned from the relatively short observational and instrumental record.

In order to attribute changes in the climate system to human influence, researchers must demonstrate that a detected change is “consistent with the estimated responses to the given combination of anthropogenic and natural forcing” and “not consistent with alternative, physically plausible explanations of recent climate change that exclude important elements of the given combinations of forcings.”¹⁰² The foundation for this analysis is physical understanding of how the climate system reacts to different radiative forcings, such as GHGs, atmospheric aerosols, solar radiation, and reflectivity (albedo), all of which influence the balance of energy in the global climate system. Scientists must also account for the global carbon cycle in order to ascertain how changes in radiative forcings will affect different components of the climate system (e.g., the relative uptake of heat and carbon dioxide by oceans). Finally, scientists must account for natural variability within the climate system in order to ascertain whether an observed change is caused by human forcing or natural variability.

The CMIP model ensembles (discussed in the section titled “Statistical Techniques and Climate Models” above) are commonly used in climate change attribution studies. Owing to ongoing advances in physical understanding, observations, and computational power, climate models now operate at finer and finer spatial scales, include interactions across more and more components of the climate system, and generate thousands of years of model output under different forcings and initial conditions. As models have grown in sophistication and complexity, their utility for climate attribution has grown—in 2014, IPCC AR5 found that models driven by historical greenhouse gas emissions and natural forcings (e.g., volcanoes and solar variability) could already “reproduce observed continental-scale surface temperature patterns and trends over many decades, including the more rapid warming since the mid-20th century and the cooling immediately following large volcanic eruptions.”¹⁰³ Models have continued to improve since then.

IPCC AR6 assessed results from climate models participating in CMIP6 and found that:

102. IPCC, *Climate Change 2001: The Scientific Basis*, Working Group I Contribution to the Third Assessment Report of the IPCC 56 (J.T. Houghton et al. eds.), <https://perma.cc/3GBS-NLYT> [hereinafter IPCC TAR WGI].

103. The IPCC issued this statement with *very high confidence*. IPCC, *Climate Change 2013: The Physical Science Basis*, Working Group I Contribution to the Fifth Assessment Report of the IPCC 15 (Thomas F. Stocker et al. eds., 2013) [hereinafter IPCC AR5 WGI], <https://perma.cc/3LKN-28W5>.

[The CMIP6 models] include new and better representations of physical, chemical and biological processes, as well as higher resolution, compared to climate models considered in previous IPCC assessment reports. This has improved the simulation of the recent mean state of most large-scale indicators of climate change and many other aspects across the climate system. Some differences from observations remain, for example in regional precipitation patterns.¹⁰⁴

However, there are still some significant differences between model results and some observations, particularly with regard to the hydrological cycle and precipitation patterns, and remaining uncertainties in cloud cover and cloud types. Global and even regional models are often too coarse in resolution to accurately and precisely simulate certain aspects of precipitation, particularly heavy precipitation (as there is significant variation in precipitation at relatively small spatial and temporal scales). This is a good example of a “downscaling” challenge in climate science that researchers are actively seeking to address.¹⁰⁵ Such downscaling challenges are most apparent in extreme-event and impact attribution, but they also appear, to a lesser extent, in climate change attribution studies. This is because many of the observed changes in the global climate system vary on a regional basis, the result of factors including differences in forcings and the higher natural variability at finer spatial scales.

Status of research

The existing body of research shows that human activities have unequivocally warmed the climate, and this has caused “[w]idespread changes in the atmosphere, ocean, cryosphere, and biosphere.”¹⁰⁶ Scientists have also made considerable

104. IPCC AR6 WGI, *supra* note 6, at 285.

105. See, e.g., Jie Chen & Xunchang John Zhang, *Challenges and Potential Solutions in Statistical Downscaling of Precipitation*, 165 *Climatic Change* 1 (2021), <https://doi.org/10.1007/s10584-021-03083-3>.

106. IPCC AR6 WGI, *supra* note 6, at 148. See also USGCRP, *Fifth National Climate Assessment*, chs. 2–4 (A.R. Crimmins et al. eds., 2023), <https://doi.org/10.7930/NCA5.2023> [hereinafter NCA5] (“[t]he evidence for warming across multiple aspects of the Earth system is incontrovertible, and the science is unequivocal that increase in atmospheric greenhouse gases are driving many observed trends and changes”). The finding that human activities have “unequivocally” caused climate change is based on the continuing upward trend in GHG emissions and related trends in climate variables such as temperature increases, ice-mass loss, and sea-level rise. Scientific confidence in this finding has increased with each subsequent IPCC assessment. See IPCC AR6 WGI, *supra* note 6, at 182 (“The Second Assessment Report (SAR) stated that ‘the balance of evidence suggests a discernible human influence on global climate.’ Five years later, the Third Assessment Report (TAR) concluded that ‘there is new and stronger evidence that most of the warming observed over the last 50 years is attributable to human activities.’ The AR4 further strengthened previous statements, concluding that ‘most of the observed increase in global average temperatures

progress toward quantifying the effect of human activities on different components of the climate system, although there is still uncertainty about some precipitation and hydrological changes attributable to climate change. There is also some uncertainty about the respective influence of different climate forcings on observed changes, including the influence of nonhuman forcings, as well as the magnitude of internal variability and its influence on observed changes. The following paragraphs provide an overview of key findings for specific variables and components within the climate system.

Chemical composition of the climate: Since the onset of the Industrial Revolution, human activities have caused significant increases in atmospheric GHG concentrations. Between 1750 and 2019, CO₂ concentrations increased 47% to 410 parts per million (ppm); CH₄ concentrations increased 156% to 1,866 parts per billion (ppb); and N₂O concentrations increased 23% to 332 ppb.¹⁰⁷ The current levels of these GHGs in the atmosphere are unprecedented on timescales spanning from thousands to millions of years.¹⁰⁸ Because of the global carbon cycle, not all GHGs remain in the atmosphere. Approximately 44% of anthropogenic CO₂ emissions have accumulated in the atmosphere, with the remainder absorbed by land and ocean CO₂ sinks.¹⁰⁹ Scientists predict that the fraction of CO₂ absorbed by land and oceans will decrease as cumulative CO₂ emissions increase.¹¹⁰

Atmospheric and surface temperature: IPCC AR6 found “unequivocal” evidence that human influence has warmed the atmosphere, oceans, and land; this warming trend is unprecedented in at least the last 2,000 years and cannot be explained by natural drivers or internal variability.¹¹¹ The USGCRP, National Academies, and other major scientific organizations have reached

since the mid-20th century is very likely due to the observed increase in anthropogenic greenhouse gas concentrations.’ The AR5 assessed that a human contribution had been detected in: changes in warming of the atmosphere and ocean; changes in the global water cycle; reductions in snow and ice; global mean sea level rise; and changes in some climate extremes. The AR5 concluded that ‘it is extremely likely that human influence has been the dominant cause of the observed warming since the mid-20th century.’”) (citations omitted).

107. IPCC AR6 WGI, *supra* note 6, at 281. This rate of change in CO₂ and CH₄ concentrations is unprecedented in at least the past 800,000 years. *Id.* at 69. GHG concentrations have continued to rise since AR6. In 2022, CO₂ concentrations reached 418 ppm, CH₄ concentrations reached 1,923 ppb, and N₂O concentrations reached 336.16 ppb. NOAA Glob. Monitoring Lab’y, *Carbon Cycle Greenhouse Gases*, <https://perma.cc/U45E-HU3R>.

108. NAS Update, *supra* note 6, at 9. IPCC AR6 found that in 2019, atmospheric CO₂ concentrations were higher than at any time in at least 2 million years (*high confidence*) and atmospheric CH₄ and N₂O concentrations were higher than at any time in at least 800,000 years (*very high confidence*). IPCC AR6 WGI, *supra* note 6, at 8.

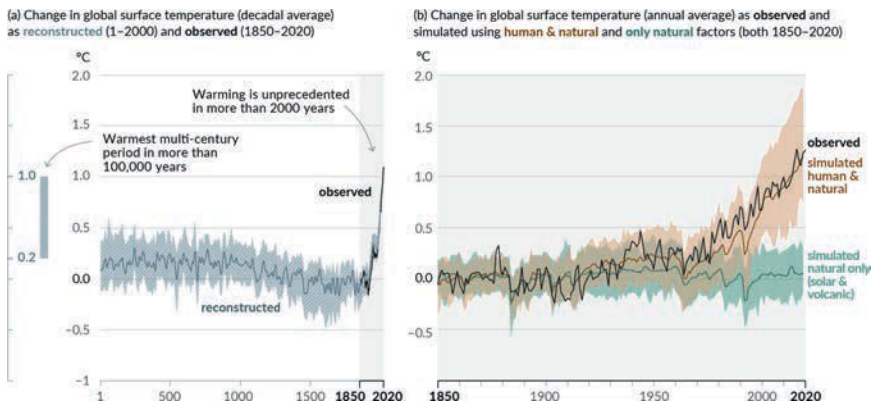
109. IPCC AR6 WGI, *supra* note 6, at 690. Specifically, the IPCC estimates that 23% of anthropogenic CO₂ emissions have been taken up by ocean sinks, and 31% have been taken up by terrestrial ecosystems (or “land sinks”). *Id.* at 80.

110. *Id.* at 744.

111. *Id.* at 4, 6.

similar conclusions.¹¹² As of 2019, the decadal average global surface temperature had increased approximately 1.09 [0.95–1.20] °C over preindustrial levels, with larger increases over land (1.59 [1.34–1.83] °C) than the ocean (0.88 [0.68–1.01] °C).¹¹³ The observed warming trend is consistent with physical understanding and model simulations of the climate forcing effects of human drivers (see Figure 1). IPCC AR6 found that it is *likely* that humans are responsible for approximately 1.07 [0.8–1.3] °C of the observed increase in global surface temperature, with well-mixed GHGs contributing to a warming of 1.0–2.0 °C and other human drivers (primarily aerosols) contributing to a cooling of 0.0–0.8 °C.¹¹⁴

Figure 1. Changes in global surface temperature relative to 1850–1900.



Panel (a) shows changes in global surface temperature reconstructed from paleoclimate archives (solid gray line, years 1–2000) and from direct observations (solid black line, 1850–2020), both relative to 1850–1900 and decadal averaged.

Source: Figure SPM.1 in IPCC, *Summary for Policymakers*, in *Climate Change 2021: The Physical Science Basis*, Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change 3–32 (Masson-Delmotte et al. eds.), <http://doi.org/10.1017/9781009157896.001>.

112. See NCA5, *supra* note 106; NAS Update, *supra* note 6.

113. IPCC AR6 WGI, *supra* note 6, at 5. The figures cited here include average estimates as well as ranges (in square brackets). These ranges reflect the 90% confidence interval (meaning that there is an estimated 90% likelihood of the value being within that range). *Id.* at 41. The IPCC AR6 uses the period of 1850–1900 as a proxy for pre-industrial temperature levels (as well as other pre-industrial climate conditions). *Id.* at 163 (see Cross-Chapter Box 11.1).

114. *Id.* at 5. Similarly, NCA4 found that there was “a likely human contribution of 93%–123% of the observed 1951–2010 change” in global temperature. NCA4 Vol. I, *supra* note 6, at 14.

The gray shading with white diagonal lines shows the *very likely* ranges for the temperature reconstructions. Panel (b) shows changes in global surface temperature over the past 170 years (black line) relative to 1850–1900 and annually averaged, compared to CMIP6 climate model simulations of the temperature response to both human and natural drivers (brown) and to only natural drivers (solar and volcanic activity, green). Solid colored lines show the multimodel average, and colored shades show the *very likely* range of simulations.

Ocean warming: As the atmosphere has warmed, so too have the oceans. IPCC AR6 found, with *high confidence*, that approximately 91% of the extra heat added to the climate system had been added to the ocean.¹¹⁵ The total ocean heat content has increased substantially, and ocean surface temperature has increased, on average, by 0.88 [0.68–1.01] °C since preindustrial times. It is extremely likely that human influence is the primary driver of ocean warming.¹¹⁶

Sea-level rise: Global sea levels have risen as a result of the melting of land-based ice and the increase in ocean heat content (since warmer waters occupy more volume from thermal expansion). Global mean sea level increased by 0.20 [0.15–0.25] meters between 1901 and 2018, and the rate of sea-level rise has been accelerating since the 1990s.¹¹⁷ It is *very likely* that human influence was the main driver of these increases, at least since 1971.¹¹⁸

Ocean deoxygenation: Ocean deoxygenation (oxygen decline) is another consequence of ocean warming (the solubility of dissolved oxygen decreases as sea water becomes warmer). There is *high confidence* that oxygen levels have dropped in many upper ocean regions since the mid-20th century, and *medium confidence* that human-induced ocean warming contributed to this drop.¹¹⁹

Ocean acidification: Researchers estimate that the surface ocean has absorbed approximately one quarter of all human CO₂ emissions,¹²⁰ causing the pH of the ocean surface to decrease. Researchers are *virtually certain* that human-caused CO₂ emissions are the main driver of ocean acidification.¹²¹

Cryosphere: There has been a substantial decline in sea ice, terrestrial glaciers, and snowpack in the past century, with considerable geographic variation in the magnitude and rate of decline. For example, IPCC AR6 concluded that (1) there has been a substantial reduction in Arctic sea ice over the period of 1979–2019, and in 2011–2022, average annual Arctic sea ice reached its lowest

115. This is consistent with the IPCC's prior assessment (AR5), in which it found that approximately 90% of the accumulated energy from human-induced climate change was stored within the oceans. See IPCC AR6 WGI, *supra* note 6, at 283 (comparing statements from AR5 and AR6).

116. *Id.* at 5, 1214.

117. *Id.*

118. *Id.*

119. *Id.* at 5, 714.

120. *Id.* at 714.

121. *Id.* at 5.

level since at least 1850 (*high confidence*); (2) there have also been substantial reductions in ice sheets: between 1992 and 2020, the Greenland ice sheet lost 4,890 [4,180–5,640] billion metric tons of mass, and the Antarctic ice sheet lost 2,670 [1,800–3,540] billion metric tons of mass; (3) glaciers lost 6,200 [4,600–7,800] billion metric tons of mass between 1993 and 2019 (*very high confidence*), and glaciers lost more mass in 2010–2019 than in any other decade since the beginning of the observational record (*very high confidence*); (4) permafrost temperatures have been increasing around the world (*high confidence*) resulting in permafrost thaw in some regions (*medium confidence*); and (5) Northern Hemisphere spring snow cover has decreased since 1978 (*very high confidence*), and Northern Hemisphere spring snow cover has likely decreased since 1950 (*high confidence*).¹²²

Hydrological cycle and precipitation: Ascertaining the effect of anthropogenic forcings on the hydrologic cycle and precipitation is one of the more challenging areas of climate change attribution,¹²³ but there is evidence that the global hydrological cycle is intensifying as a result of anthropogenic climate change. IPCC AR6 expressed *high confidence* that climate change has contributed to an overall increase in atmospheric moisture content (i.e., water vapor) as well as precipitation intensity.¹²⁴ Scientists also expect that global warming will cause an overall increase in global precipitation,¹²⁵ but this “has not yet been detected and attributed to human activities given large observational uncertainties and low signal-to-noise ratio.”¹²⁶ There is evidence that climate change is causing regional increases or decreases in average precipitation, with some areas becoming wetter and others becoming more arid.¹²⁷ Increases in the frequency and intensity of heavy precipitation events have also been attributed to human influence.¹²⁸

122. *Id.* at 1215–16.

123. Part of the challenge is detecting change—spatial gradients of precipitation can be quite large in some regions, and historical rainfall records are incomplete and contain mixed findings about the extent to which precipitation patterns have (or have not) changed since the preindustrial era. It is also difficult to attribute precipitation changes to human influence on climate because precipitation is characterized by large natural variability across a range of timescales, ranging from the intra-annual to the centennial.

124. IPCC AR6 WGI, *supra* note 6, at 1057, 1080–81. It is also *extremely likely* that climate change has driven more evaporation over the oceans, which contributes to increased water vapor as well as heavier precipitation. *Id.* at 1080.

125. *Id.* § 8.2.1.

126. *Id.* at 1079.

127. *Id.* at 1057. *See also* WMO (2022), *supra* note 62 (finding that many of the world’s regions experienced above-normal precipitation in 2022, without specifying the extent to which this trend is attributable to anthropogenic climate forcing).

128. IPCC AR6 WGI, *supra* note 6, at 8.

Extreme Event Detection and Attribution

There are multiple pathways through which climate change can influence the intensity, duration, frequency, and other characteristics of extreme events. A warmer climate increases the probability and frequency of very hot days and nights, thus contributing to heat waves. A warmer climate also contributes to increased land evaporation and drier conditions, potentially increasing the severity of droughts and wildfire conditions. The intensification of the global hydrological cycle and increases in atmospheric moisture content contribute to more severe precipitation events. Increases in atmospheric moisture content and heat, coupled with increases in sea-surface and ocean temperature, also contribute to more severe tropical cyclones. All things being equal, more heat in the climate system corresponds with more energy to power storms, since atmospheric heat can be converted into kinetic energy (e.g., in the form of high winds).

Methods and parameters

An “extreme” weather or climate event is defined by the IPCC as “the occurrence of a value of a weather or climate variable above (or below) a threshold value near the upper (or lower) ends (tails) of the range of observed values of the variable.”¹²⁹ Attribution of extremes tends to be more challenging than attribution of means for several reasons: (1) the local nature and short duration of any extremes makes them more difficult to model given the coarse resolution of climate models, (2) the relative rarity of extreme events at a given location makes it more difficult to detect and attribute a climate change “signal” amid the large “noise” of internal variability, and (3) the causal chains for extremes can be quite complicated (e.g., highly nonlinear).¹³⁰ There are also some modeling challenges relevant to extreme-event attribution. For example, models may be too Gaussian in their extreme events (i.e., they don’t produce enough of them).¹³¹ Scientists

129. *Id.* at 111.

130. Sebastian Sippel et al., *Warm Winter, Wet Spring, and an Extreme Response in Ecosystem Functioning on the Iberian Peninsula*, in BAMS 2016, *supra* note 101, at S80, <https://doi.org/10.1175/BAMS-D-17-0135.1> (citing Dorothea Frank et al., *Effects of Climate Extremes on the Terrestrial Carbon Cycle: Concepts, Processes and Potential Future Impacts*, 21 *Glob. Climate Change Biology* 2861 (2015), <https://doi.org/10.1111/gcb.12916>, and John A. Arnone III et al., *Prolonged Suppression of Ecosystem Carbon Dioxide Uptake After an Anomalously Warm Year*, 455 *Nature* 383 (2008), <https://doi.org/10.1038/nature07296>).

131. Bo Christiansen, *The Role of the Selection Problem and Non-Gaussianity in Attribution of Single Events to Climate Change*, 28 *J. Climate* 9873 (2015), <https://doi.org/10.1175/JCLI-D-15-0318.1>. “Gaussian” refers to having the shape of a normal curve or normal distribution, and extreme events may not adhere to a Gaussian distribution. A normal distribution is characterized by a relatively large frequency of values near the mean, with a relatively small frequency of values far from the mean.

have devised statistical approaches to avoid the problems and limitations associated with climate models, but these approaches rely on simplifying assumptions.¹³² Despite these complicating factors, extreme-event studies can generate reliable results for many extreme events, though confidence in attribution findings depends on the type of extreme and the location.

The results of extreme-event studies are sensitive to how the research question is framed, and what methodological approaches and datasets are used. Some studies focus on individual events; others deal with a class of events over a time frame (e.g., the 2020 North Atlantic hurricane season).¹³³ One critical framing question is how to define the event parameters for the purposes of the study. These parameters include the physical threshold for the extreme event (e.g., a certain temperature threshold for a heat wave) as well as the time frame and spatial scale for the study. There are often multiple metrics that could be used for any given event—for example, a heat wave could be defined based on absolute maximum temperature, average maximum temperature over a duration of time, or a combination of temperature and humidity. These framing decisions can affect the results of the study. For example, if researchers were to use the maximum temperature during a heat wave as the temperature threshold, and then focus their analysis on the location that reached the highest temperature during the event, the heat wave may appear more exceptional (or “extreme”) than if the temperature and spatial scale were selected in a more generic way. This is an example of how event framing could introduce selection bias into an attribution study. This is not an insurmountable obstacle—for example, efforts are underway to standardize how extreme events are defined and selected for analysis, and this would have the added benefit of facilitating more systematic comparison between extreme-event studies.¹³⁴

As discussed earlier in this reference guide, there are also different approaches to attribution.¹³⁵ Some studies use a probabilistic or risk-based approach, examining whether and to what extent climate change increased the probability (or risk) of an extreme event. Other studies use a storyline or mechanistic approach, examining how a well-understood aspect of climate change may have affected

132. NASEM (2016), *supra* note 91; Christiansen, *supra* note 131.

133. See, e.g., Kevin A. Reed et al., *Attribution of 2020 Hurricane Season Extreme Rainfall to Human-Induced Climate Change*, 13 *Nature Commc'ns* 1 (2022), <https://doi.org/10.1038/s41467-022-29379-1>.

134. NASEM (2016), *supra* note 91.

135. Although many extreme-event attribution studies clearly fall within one of these two categories, there are also some other approaches in the literature. As one example, Mann et al. suggested a modification to traditional frequentist statistical inference approach, that builds in prior physical understanding and updates based on experience. Michael E. Mann et al., *Assessing Climate Change Impacts on Extreme Weather Events: The Case for an Alternative (Bayesian) Approach*, 144 *Climatic Change* 131 (2017), <https://doi.org/10.1007/s10584-017-2048-3>.

the magnitude and/or other characteristics of an extreme event. Both approaches have benefits and drawbacks.

In probabilistic studies, researchers evaluate the likelihood of a narrowly defined climatological extreme (e.g., a temperature of 95°F) occurring with and without human influence on the climate system. This is typically accomplished by using climate models to run two types of simulations: one that represents the world as it is and another that represents a hypothetical world without anthropogenic climate forcing. Researchers will then evaluate the probability of the extreme occurring in both simulations. The results are often expressed in terms of the relative risk of the extreme occurring with and without human influence on climate, which may be expressed using a risk ratio, which is the ratio of the probability of an event occurring with and without climate change,¹³⁶ or fraction of attributable risk (FAR). In mathematical terms:¹³⁷

$$\text{FAR} = 1 - P_0/P_1$$

P_1 equals the probability of a climatic event (such as a heat wave) occurring in the presence of anthropogenic forcing of the climate system, and P_0 equals the probability of the event occurring if the anthropogenic forcing were not present. If FAR equals zero, it means that anthropogenic climate change had no effect on the probability of the event occurring; if FAR equals one, it means that the event could not have happened in the absence of anthropogenic climate change; if FAR equals 0.5, it means that anthropogenic climate change doubled the probability of the event occurring. In multi-event studies, a FAR of 0.5 can be interpreted as meaning that half of the events would not have happened in a world without anthropogenic climate change.¹³⁸

This approach was pioneered by Myles Allen in a 2003 study in which he introduced the concept of FAR as a potential basis for liability for climate damages.¹³⁹ Many other studies have since replicated Allen's approach, estimating the FAR for a range of extreme events including heat waves, droughts, and floods. This methodology derives from common approaches used in epidemiological

136. Multiple risk ratios may also be used to express probabilities of occurrence at varying levels of climate change or global warming. See, e.g., V.V. Kharin et al., *Risks from Climate Extremes Change Differently from 1.5°C to 2.0° Depending on Rarity*, 6 *Earth's Future* 704 (2018), <https://doi.org/10.1002/2018EF000813>.

137. IPCC, *Climate Change 2007*, *supra* note 48; Myles Allen, *Liability for Climate Change*, 421 *Nature* 891 (2003), <https://doi.org/10.1038/421891a>.

138. While the term FAR is typically used in extreme-event attribution, probabilistic analysis is prevalent across all forms of attribution, and the concept of attributable risk can in principle be applied to both mean changes in climate and a variety of climate impacts. See, e.g., Thomas Knutson et al., *CMIP5 Model-Based Assessment of Anthropogenic Influence on Record Global Warmth During 2016*, in *BAMS 2016*, *supra* note 101, at S11.

139. Allen, *supra* note 137.

studies and other risk-focused research. The advantages of this approach are that it is relatively well established, understood, and accepted by the scientific community,¹⁴⁰ and it provides quantitative (probabilistic) findings similar to those that are often dealt with by policy makers, planners, and courts. Drawbacks include: (1) overreliance on climate models, which as noted earlier may not be able to simulate some types of extremes with fidelity in a baseline climate, and could have blind spots with respect to how climate change may be modifying key processes influencing the extreme event, and (2) susceptibility to Type II errors (i.e., false negatives), where the signal-to-noise ratio for an event is small because of large internal variability of the atmosphere, which is often the case for dynamically driven events such as extreme precipitation and storms.¹⁴¹ Thus the probabilistic approach, like all approaches, is an imperfect measure of the precise degree to which anthropogenic influence has increased the likelihood of an event.

The mechanistic or storyline approach, introduced by Trenberth et al. (2015),¹⁴² focuses on how anthropogenic forcing may have modified the characteristics of a given event, like a regional heat wave. This involves reconstructing the causal chain of events that resulted in the extreme event, typically focusing on a few central causal factors to establish attribution.¹⁴³ Mechanistic studies tend to focus on the relationship between the observed event and well-understood components of climate change, such as warming temperatures and thermodynamics. Because of this, mechanistic studies can generate higher confidence statements with regard to those specific components. The results of this analysis may be quantitative or qualitative—for example, “warming of the upper ocean and atmosphere enabled more rainfall during event Y than otherwise would have occurred” or “caused a 30% increase in rainfall over what would have occurred.”

As with the probabilistic approach, there are advantages and drawbacks to the mechanistic approach. The chief advantages are that the mechanistic approach provides additional insights on how climate change is affecting physical characteristics of extreme events, and it allows for higher confidence attribution regarding the influence of certain aspects of climate change on certain types of events. But there are also criticisms of this approach, as articulated by Otto (2016).¹⁴⁴ First, mechanistic studies may not have the same broad relevance as probabilistic studies because they focus on how climate change has influenced

140. See NASEM (2016), *supra* note 91, at 3.

141. Kevin E. Trenberth et al., *Attribution of Climate Extreme Events*, 5 *Nature Climate Change* 725 (2015), <https://doi.org/10.1038/nclimate2657>.

142. *Id.*

143. Lloyd & Shepherd, *supra* note 71; Linda van Garderen et al., *A Methodology for Attributing the Role of Climate Change in Extreme Events: A Global Spectrally Nudged Storyline*, 21 *Nat. Hazards & Earth Sys. Scis.* 171 (2021), <https://doi.org/10.5194/nhess-21-171-2021>.

144. Friederike E.L. Otto et al., *The Attribution Question*, 6 *Nature Climate Change* 813 (2016), <https://doi.org/10.1038/nclimate3089>.

the specific characteristics of a particular event (whereas the findings from probabilistic studies are more easily applied to a class of events—e.g., all heat waves with temperatures above a certain threshold in a region). Second, it is an oversimplification to assume that some aspects of the climate system (and climate change) are well understood and others are not—rather, there is a gradient of understanding across the various components of the climate system, and that understanding is constantly evolving with new research.¹⁴⁵ Neglecting certain aspects of or processes within the climate system can render these studies incomplete.¹⁴⁶

While there is some debate about the relative merits of these two approaches, the reality is that they are complementary—they each provide different insights on the effect of anthropogenic climate change on extreme events, and one approach can be used to fill gaps where the other is unsuitable. For example, the probabilistic/risk-based approach may be more justifiable for analyzing all events above or below a certain threshold, for a class of events that are relatively well simulated by climate models (e.g., temperature extremes), whereas the storyline approach may be more appropriate for complex, iconic, multivariate events.¹⁴⁷

Status of research

In early IPCC assessments, scientists recognized that climate change would affect the frequency, intensity, spatial extent, duration and timing of weather and climate extremes, but there was low confidence in the attribution of specific extreme events to human influence on climate.¹⁴⁸ Since then, scientific

145. For example, Michael E. Mann notes that dynamical changes with warming are starting to come into focus: more specifically, a growing body of work based on observations and simple models supports the idea that the latitudinal pattern of mean temperature changes (including Arctic amplification) may facilitate changes in atmospheric dynamics that increase wave resonance and “stuck” weather, which enhances the magnitude and duration of extremes. It should be noted that global climate models generally do not reproduce this pattern of wave resonance and “stuck” weather with warming. Michael E. Mann et al., *Influence of Anthropogenic Climate Change on Planetary Wave Resonance and Extreme Weather Events*, 7 *Sci. Reps.* 45242 (2017), <https://doi.org/10.1038/srep45242>.

146. More specifically, focusing on thermodynamics—but not dynamics—within the climate system can result in an incomplete analysis, in part because of the failure to capture interactions between thermodynamics and dynamics. Otto shows how the dynamics and thermodynamics counteracted each other in 2013 German floods. See Otto et al., *supra* note 144, at 815. Similarly, a study in Western Australia found dynamics/circulation changes that favor less rain, but thermodynamic (specifically sea surface temperature) changes that favor increase in rain. Thomas L. Delworth & Fanrong Zeng, *Regional Rainfall Decline in Australia Attributed to Anthropogenic Greenhouse Gases and Ozone Levels*, 7 *Nature Geoscience* 583 (2014), <https://doi.org/10.1038/ngeo2201>.

147. Elisabeth A. Lloyd & Naomi Oreskes, *Climate Change Attribution: When Is It Appropriate to Accept New Methods?*, 6 *Earth's Future* 311 (2018), <https://doi.org/10.1002/2017EF000665>.

148. See IPCC TAR WGI, *supra* note 102, at ch. 3.

confidence in extreme-event attribution has advanced significantly as a result of “better physical understanding of processes, an increasing proportion of the scientific literature combining different lines of evidence, and improved accessibility to different types of climate models (*high confidence*).”¹⁴⁹ IPCC AR6 noted that the evidence of observed changes in extremes such as heat waves, heavy precipitation, droughts, and tropical cyclones, and their attribution to climate change, had strengthened since AR5.¹⁵⁰ According to IPCC AR6, it is “now an established fact that human-induced greenhouse gas emissions have led to an increased frequency and/or intensity of some weather and climate extremes,” particularly temperature extremes.¹⁵¹

The following paragraphs provide a summary of recent attribution findings for different classes of extreme events, drawing primarily on IPCC AR6 and findings from the U.S. Fifth National Climate Assessment (NCA5).

Extreme heat: An increase in the magnitude, frequency, and duration of extreme heat events is a direct and foreseeable consequence of a warming climate.¹⁵² NCA5 concluded, with *very high confidence*, that the frequency and intensity of extreme heat events are increasing in most continental regions of the world, consistent with expected physical responses to a warming climate.¹⁵³ IPCC AR6 similarly found that it is “*virtually certain* that hot extremes (including heatwaves) have become more frequent and more intense across most land regions since the 1950s,” and there is “*high confidence* that human-induced climate change is the main driver of these changes.”¹⁵⁴ Moreover, “[s]ome recent hot extremes observed over the past decade would have been *extremely unlikely* to occur without human influence on the climate system.”¹⁵⁵ Importantly, climate change is not only contributing to extreme heat events over land—according to IPCC AR6, marine heat waves have “approximately doubled in frequency since the 1980s (*high confidence*) and human influence has *very likely* contributed to most of them since at least 2006.”¹⁵⁶ The USGCRP and IPCC findings are consistent with a growing number of studies on extreme heat and climate change, many of which have demonstrated a very strong anthropogenic signal in such events.¹⁵⁷

149. IPCC AR6 WGI, *supra* note 6, at 1517. See also NASEM (2016), *supra* note 91 (discussing advances in extreme-event attribution).

150. IPCC, AR6 Synthesis Report: Climate Change 2023, <https://perma.cc/PBK3-NMUD>.

151. IPCC AR6 WGI, *supra* note 6, at 1517.

152. These core characteristics of extreme heat events (magnitude, duration, frequency) are all highly sensitive to changes in mean temperatures. Radley M. Horton et al., *A Review of Recent Advances in Research on Extreme Heat Events*, 2 Current Climate Change Reps. 242 (2016), <https://doi.org/10.1007/s40641-016-0042-x>.

153. NCA5, *supra* note 106, at 2–38; NCA4 Vol. I, *supra* note 6, at 19.

154. IPCC AR6 WGI, *supra* note 6, at 8.

155. *Id.*

156. *Id.*

157. See, e.g., Peter Stott et al., *Human Contribution to the European Heatwave of 2003*, 432 Nature 610 (2004), <https://doi.org/10.1038/nature03089>; Noah S. Diffenbaugh et al., *Quantifying*

Droughts and aridity: Although warmer temperatures are one relatively well-understood factor that can contribute to droughts, it is more challenging to isolate the effect of anthropogenic climate change on dryness and drought conditions. Droughts are highly complex meteorological events with many different factors affecting their likelihood, severity, duration, and other characteristics, and there is significant natural variability in precipitation, which makes it difficult to identify the anthropogenic signal through the noise of variability.¹⁵⁸ However, researchers have identified evidence of an anthropogenic signal in the heat-related aspects of drought (e.g., increased evapotranspiration) and have been able to estimate the extent to which warmer temperatures have intensified drought conditions.¹⁵⁹ Based on this research, IPCC AR6 expressed *medium confidence* that human-induced climate change had contributed to ecological and agricultural droughts in some regions due to increased land evapotranspiration.¹⁶⁰ IPCC AR6 did not find clear evidence that anthropogenic climate change was causing an increase in meteorological or hydrological droughts in most regions

the Influence of Global Warming on Unprecedented Extreme Climate Events, 114 PNAS 4881 (2017), <https://doi.org/10.1073/pnas.1618082114>; Yukiko Imada et al., *Climate Change Increased the Likelihood of the 2016 Heat Extremes in Asia*, in BAMS 2016, *supra* note 101, at S97, <https://doi.org/10.1175/BAMS-D-17-0109.1>; John Walsh et al., *The High Latitude Marine Heat Wave of 2016 and Its Impacts on Alaska*, in BAMS 2016, *supra* note 101, at S39, <https://doi.org/10.1175/BAMS-D-17-0105.1>; S.E. Perkins-Kirkpatrick et al., *The Role of Natural Variability and Anthropogenic Climate Change in the 2017/18 Tasman Sea Marine Heatwave*, 100 Bull. Am. Meteorological Soc'y S105 (2019), <https://doi.org/10.1175/BAMS-D-18-0116.1>; Alexander Robinson et al., *Increasing Heat and Rainfall Extremes Now Far Outside the Historical Climate*, 4 NPJ Climate & Atmospheric Sci. 45 (2021), <https://doi.org/10.1038/s41612-021-00202-w>.

158. There are several different types of droughts recognized in the literature, including: (1) meteorological droughts, which are defined by the degree and duration of dryness or rainfall deficit (in comparison to a normal or average amount); (2) hydrological droughts, which are defined by the impact of dryness and rainfall deficits on water supplies such as lakes, rivers, streams, reservoirs, and aquifers; (3) agricultural droughts, which are defined by the effect of dryness and rainfall deficits on agricultural systems; and (4) ecological droughts, which are defined by the effect of dryness and rainfall deficits on ecosystems. Attribution for hydrological droughts is particularly challenging because of the complexity of these events, limited data about historical trends, and the fact that human water consumption and management decisions are leading mechanisms behind such droughts (potentially making it harder to detect a climate change signal). See IPCC AR6 WGI, *supra* note 6, at 1576–78.

159. See, e.g., T.R. Marthews et al., *The 2014 Drought in the Horn of Africa: Attribution of Meteorological Drivers*, in *Explaining Extreme Events of 2014 from a Climate Perspective*, 96 Bull. Am. Meteorological Soc'y S1, S83 (2015), <https://doi.org/10.1175/BAMS-D-15-00115.1> [hereinafter BAMS 2014]. See also Eduardo S.P.R. Martins et al., *A Multimethod Attribution Analysis of the Prolonged Northeast Brazil Hydrometeorological Drought (2012–16)*, in BAMS 2016, *supra* note 101, at S65, <https://doi.org/10.1175/BAMS-D-17-0102.1>; Xing Yuan et al., *Anthropogenic Intensification of Southern African Flash Droughts as Exemplified by the 2015/16 Season*, in BAMS 2016, *supra* note 101, at S86, <https://doi.org/10.1175/BAMS-D-17-0077.1>; Chris Funk et al., *Anthropogenic Enhancement of Moderate-to-Strong El Niño Events Likely Contributed to Drought and Poor Harvests in Southern Africa During 2016*, in BAMS 2016, *supra* note 101, at S91, <https://doi.org/10.1175/BAMS-D-17-0112.1>.

160. IPCC AR6 WGI, *supra* note 6, at 8.

of the world; however, it did express *medium confidence* that precipitation deficits had increased in several regions across all continents.¹⁶¹

Extreme precipitation: The IPCC and NCAs have both found clear evidence that heavy rainfall events are increasing around the world and in the United States, and this is generally consistent with expected physical responses to a warming climate.¹⁶² IPCC AR6 specifically found that the “frequency and intensity of heavy precipitation events have increased since the 1950s over most land areas for which observational data are sufficient for trend analysis (*high confidence*), and human-induced climate change is *likely* the main driver.”¹⁶³ This is another area where there have been significant advances in detection and attribution research: whereas IPCC AR5 only expressed *medium confidence* in the attribution of extreme precipitation events, IPCC AR6 found that there was “new and *robust evidence* of human influence on extreme precipitation.”¹⁶⁴

The IPCC’s assessment relates to an overall increase in extreme precipitation at a global scale, but there are important regional and seasonal variations in rainfall. The dynamic nature of extreme precipitation events and the large internal variability in precipitation can make it more difficult to attribute specific events to climate change. The storyline approach to attribution was developed in part to improve attribution for difficult-to-model events like extreme precipitation. Researchers used this approach to examine the effect of anthropogenic climate change on the 2013 floods in Boulder, Colorado, and found that anthropogenic drivers increased the magnitude of the rainfall for that week by approximately 30%.¹⁶⁵ The scientists also conducted a probabilistic analysis of potential impacts on flooding and found that this 30% increase in rainfall approximately doubled the likelihood of flood-inducing rainfall occurring during that event. In contrast, researchers using the probabilistic approach to attribution of the Boulder floods found no evidence that anthropogenic climate change had increased the probability of the event occurring.¹⁶⁶ This underscores the sensitivity of results to methodological choices made in extreme-event attribution.

161. *Id.*

162. See NCA4 Vol. I, *supra* note 6, at 19; IPCC AR6 WGI, *supra* note 6, at 8. The USGCRP has also found “robust evidence that human-caused warming has contributed to increases in the frequency and severity of the heaviest precipitation events across nearly 70% of the [United States].” NCA5, *supra* note 106, at 2–18.

163. IPCC AR6 WGI, *supra* note 6, at 8.

164. *Id.* at 1562.

165. Pardeep Pall et al., *Diagnosing Conditional Anthropogenic Contributions to Heavy Colorado Rainfall in September 2013*, 17 *Weather & Climate Extremes* 1 (2017), <https://doi.org/10.1016/j.wace.2017.03.004>.

166. See Martin Hoerling et al., *Northeast Colorado Extreme Rains Interpreted in a Climate Change Context*, 95 *Bull. Am. Meteorological Soc’y* S1, S15 (2014), <https://perma.cc/6BLT-BXTZ>.

Tropical and extratropical cyclones: Physical understanding suggests that tropical cyclones will become more severe in a warmer climate.¹⁶⁷ Cyclones derive energy from sea surface temperature, and warmer sea surface temperatures, all other things being equal, will increase the intensity of storms (in terms of wind speed, precipitation, and storm surge).¹⁶⁸ A warmer ocean also produces more evaporation, and a warmer atmosphere holds more moisture, thus contributing to heavier rainfall and flooding. IPCC AR6 expressed *high confidence* that anthropogenic climate change had contributed to increases in heavy precipitation associated with tropical cyclones¹⁶⁹ and that it was *likely* that the global portion of major (Category 3–5) tropical occurrences had increased over the past four decades.¹⁷⁰ There is more uncertainty about the effect of a warming climate on extratropical cyclone activity.¹⁷¹ Few attribution studies have found a discernible anthropogenic influence on extratropical cyclones because of factors such as large interannual-to-decadal variability in such cyclones,¹⁷² and thus the IPCC has expressed *low confidence* in the detection and attribution of changes in extratropical cyclones.¹⁷³ However, since research was compiled for IPCC AR6, there have been some studies finding that anthropogenic forcings likely have contributed to more severe extratropical cyclones.¹⁷⁴

Compound extremes: Compound extreme events are those that occur because of the interaction of multiple climate-related extremes. While multiple typologies of compound extremes have been identified, including sequences of events, and simultaneous events in multiple regions,¹⁷⁵ we focus here on the most studied category: multivariate (multivariable) extreme events at a single time and location. Examples include fire weather conditions (a combination of hot, dry, and windy conditions), compound flooding (e.g., flooding caused by a

167. See IPCC AR6 WGI, *supra* note 6, at 70 (“The proportion of tropical cyclones that are intense is expected to increase (*high confidence*), but the total global number of tropical cyclones is expected to decrease or remain unchanged (*medium confidence*).”). Some of the other factors that can influence tropical cyclones include aerosols and dust, wind shear, temperatures in the upper atmosphere, and wave disturbances.

168. See, e.g., Sally L. Lavender et al., *The Influence of Sea Surface Temperature on the Intensity and Associated Storm Surge of Tropical Cyclone Yasi: A Sensitivity Study*, 18 Nat. Hazards & Earth Sys. Scis. 795 (2018), <https://doi.org/10.5194/nhess-18-795-2018>.

169. IPCC AR6 Synthesis Report, *supra* note 150, at 51; IPCC AR6 WGI, *supra* note 6, at 67 (Table TS.2).

170. IPCC AR6 WGI, *supra* note 6, at 9, 67 (Table TS.2).

171. *Id.* at 70.

172. See, e.g., Frauke Feser et al., *Hurricane Gonzalo and Its Extratropical Transition to a Strong European Storm*, in BAMS 2014, *supra* note 159, at S51, <https://doi.org/10.1175/BAMS-D-15-00122.1>.

173. IPCC AR6 WGI, *supra* note 6, at 70, 338.

174. See, e.g., Mireia Ginesta et al., *A Methodology for Attributing Severe Extratropical Cyclones to Climate Change Based on Reanalysis Data: The Case Study of Storm Alex 2020*, 61 Climate Dynamics 229 (2023), <https://doi.org/10.1007/s00382-022-06565-x>.

175. Jakob Zscheischler et al., *A Typology of Compound Weather and Climate Events*, 1 Nature Revs. Earth & Env’t 333 (2020), <https://doi.org/10.1038/s43017-020-0060-z>.

combination of storm surge and heavy rainfall, or rapid snowmelt and heavy rainfall), and concurrent heat waves and droughts.¹⁷⁶ Scientific understanding of anthropogenic influence on such events largely depends on understanding of the various component events. IPCC AR6 found that human influence has *likely* increased the probability of various compound extreme events since the 1950s.¹⁷⁷ There is *high confidence* that anthropogenic climate change has increased the frequency of concurrent heat waves and droughts on a global scale and *medium confidence* that it has increased the frequency of fire weather and compound flooding in some regions.¹⁷⁸ There is also *high confidence* that regions and connected sectors will experience multiple regional extreme events at the same time,¹⁷⁹ placing greater stress on both human and natural systems.¹⁸⁰

Impact Detection and Attribution

Many of the processes and phenomena described in the sections above could be characterized as “impacts” of climate change (e.g., sea-level rise, more severe extreme events). However, for the purposes of this reference guide, we use the term “impact attribution” to describe research on the effects of climate change on humans and ecosystems. This is consistent with the approach taken in recent IPCC assessments, specifically the division between Working Group I (WGI), which synthesizes research on the physical science basis for anthropogenic climate change, and Working Group II (WGII), which synthesizes research on the impacts of climate change.¹⁸¹

176. See IPCC AR6 WGI, *supra* note 6, at 9 n.18.

177. *Id.* at 9.

178. *Id.* Research published after AR6 has identified a strong causal association between climate change and increases in fire weather in some regions. See, e.g., Zhongwei Liu et al., *The April 2021 Cape Town Wildfire: Has Anthropogenic Climate Change Altered the Likelihood of Extreme Fire Weather?*, 104 Bull. Am. Meteorological Soc’y E298 (2023), <https://doi.org/10.1175/BAMS-D-22-0204.1>; Michael Goss et al., *Climate Change Is Increasing the Likelihood of Extreme Autumn Wildfire Conditions Across California*, 15 Env’t Rsch. Letters 094016 (2020), <https://doi.org/10.1088/1748-9326/ab83a7>; Simon F.B. Tett et al., *Anthropogenic Forcings and Associated Changes in Fire Risk in Western North America and Australia During 2015/16*, in BAMS 2016, *supra* note 101, at S60, <https://doi.org/10.1175/BAMS-D-17-0096.1>.

179. IPCC AR6 WGI, *supra* note 6, at 135.

180. See section titled “Impact Detection and Attribution.”

181. See IPCC, Climate Change 2022: Impacts, Adaptation, and Vulnerability, Working Group II Contribution to the Sixth Assessment Report of the IPCC 2912 (Hans Otto Pörtner, et al. eds., 2022), <https://perma.cc/VL7D-TMET> [hereinafter IPCC AR6 WGII] (the term “impacts” refers to the effects of climate change on “natural and human systems” including effects on “lives, livelihoods, health and well-being, ecosystems and species, economic, social and cultural assets, services (including ecosystem services), and infrastructure”).

Impact attribution deals with the consequences and outcomes that are most relevant in legal discourse and litigation—specifically, the question of who will be harmed by climate change and to what extent.¹⁸² Impact attribution deals with consequences that are further along the causal chain, and this adds a layer of complexity to the attribution analysis. Nonetheless, there are many impacts that have been attributed to anthropogenic climate forcing with a high level of confidence and certainty.

Methods and parameters

As with other areas of detection and attribution, researchers use physical understanding, observational data, statistical analysis, and climate models to detect and attribute impacts to anthropogenic climate forcing. However, as discussed earlier in this reference guide, there are some unique challenges associated with impact attribution.¹⁸³ In particular, impact attribution researchers must account for a larger number of exogenous variables and processes (i.e., those not related to the climate system) that influence the impact being studied. For example, in order to attribute monetary damages or human casualties from an extreme event to anthropogenic climate change, researchers must account for other factors that contributed to those damages. Data on these exogenous variables and processes may be limited, thus contributing to uncertainty about the relative roles of climate change and other drivers in explaining an observed impact. Furthermore, in most cases there is not a linear cause-and-effect relationship between changes in the climate system and specific impacts on humans and ecosystems. Each additional degree of warming may cause far more damage than the previous degree of warming, and climate change impacts on human and ecosystems may feed back on the climate in complex ways (for example, irrigation of crops in response to extreme temperature may itself impact temperature). Some of the challenges associated with extreme-event attribution also apply to impact attribution (potentially to an even greater extent)—for example, the spatial and temporal scale of an impact may be too fine to capture with existing climate or sectoral impact models.

Nonetheless, researchers can draw fairly robust conclusions about the general causal connection between climate change and many types of impacts, and in some cases, it is possible to quantify the contribution of anthropogenic forcing to specific damages, harms, and economic and noneconomic losses.¹⁸⁴ The use of fixed-effect regression methods to infer causation in large datasets, initially for

182. See Burger et al., *supra* note 77, at 111.

183. See section titled “Challenges Associated with Downscaling and Exogenous Variables” above.

184. IPCC AR6 WGII, *supra* note 181, at 8.

microeconomic applications, has been particularly noteworthy.¹⁸⁵ For example, researchers have developed techniques for estimating the proportion of monetary damages incurred during an extreme event that is attributable to climate change¹⁸⁶ and the public health impacts attributable to climate change.¹⁸⁷ Qualitative analyses can also provide valuable insights, including explanations of the processes and mechanisms by which climate change is affecting a particular system or outcome.¹⁸⁸ Qualitative analyses can be used where there are exogenous and confounding variables that would impede precise quantification of impacts associated with climate change.

Status of research

In recent assessments, the IPCC has found increasingly robust evidence of substantial and wide-ranging impacts of climate change across all climate zones and continents. In particular, observed increases in the frequency and intensity of climate and weather extremes have caused “widespread, pervasive impacts to ecosystems, people, settlements, and infrastructure” (*high confidence*).¹⁸⁹ Slow-onset processes, such as increases in atmospheric and ocean temperatures, ocean acidification, sea-level rise, and regional decreases in average precipitation have also affected human and natural systems across the world (*high confidence*).¹⁹⁰ Some of the impacts that have been attributed to anthropogenic climate forcing include increases in heat-related human mortality (*medium confidence*), coral bleaching and mortality (*high confidence*), increased drought-related tree mortality (*high confidence*), increases in areas burned by wildfires (*medium to high confidence*, depending on the region), and increases in storm-related losses and damages due to sea-level rise and increases in heavy precipitation (*medium confidence*).¹⁹¹ These impacts are disproportionately affecting “the most vulnerable people and

185. See, e.g., Christopher W. Callahan & Justin S. Mankin, *National Attribution of Historical Climate Damages*, 172 *Climatic Change* 1 (2022), <https://doi.org/10.1007/s10584-022-03387-y>. For a broader review of climate econometrics, see Solomon Hsiang, *Climate Econometrics*, 8 *Ann. Rev. Res. Econ.* 43 (2016), <https://perma.cc/EG9G-G7ZA>.

186. See, e.g., Benjamin H. Strauss et al., *Economic Damages from Hurricane Sandy Attributable to Sea Level Rise Caused by Anthropogenic Climate Change*, 12 *Nature Commc'ns* 2720 (2021), <https://doi.org/10.1038/s41467-021-22838-1>; Frame et al., *supra* note 101.

187. See, e.g., Kristie L. Ebi et al., *Using Detection and Attribution to Quantify How Climate Change Is Affecting Health*, 39 *Health Affs.* 2168 (2020), <https://doi.org/10.1377/hlthaff.2020.01004>.

188. See, e.g., Tom H. Oliver & Mike D. Morecroft, *Interactions Between Climate Change and Land Use Change on Biodiversity: Attribution Problems, Risks, and Opportunities*, 5 *WIREs Climate Change* 317 (2014), <https://doi.org/10.1002/wcc.271>.

189. IPCC AR6 WGII, *supra* note 181, at 9.

190. *Id.*

191. *Id.*

systems” across different regions, and some natural and human systems have been “pushed beyond their ability to adapt” (*high confidence*).¹⁹²

Ecosystems, species, and ecological indicators: There is robust evidence that climate change is adversely affecting ecosystems and species in every region of the world, and the extent and magnitude of these impacts are larger than estimated in previous assessments. IPCC AR6 found, with *high confidence*, that climate change has caused “substantial damage, and increasingly irreversible losses, in terrestrial, freshwater and coastal and open marine ecosystems.”¹⁹³ Some ecosystems are already “reaching or surpassing hard adaptation limits,” including warm-water coral reefs, some coastal wetlands, some rainforests, and some polar and mountain ecosystems (*high confidence*).¹⁹⁴ Climate change is also leading to changes in the geographic distribution of some species and the spread of insect pests and invasive species. Climate-change-fueled ecosystem deterioration is having “adverse socioeconomic consequences” on the communities and sectors that depend on these ecosystems (*high confidence*).¹⁹⁵ For example, impacts on marine ecosystems affect fisheries’ health and productivity, and phenological changes associated with climate change, such as longer growing seasons, affect agriculture and food-production systems.

Physical impacts of extreme weather: Many attribution studies straddle the line between extreme-event and impact attribution, examining the effect of anthropogenic forcing on climatological extremes (e.g., heavy precipitation) and corresponding physical hazards (e.g., floods). In some cases, the effect of climate change on physical hazards is quite clear. For example, IPCC AR6 expressed *high confidence* that sea-level rise has contributed to increased coastal flooding in low-lying areas.¹⁹⁶ However, there is less certainty regarding how climate change is affecting inland flooding conditions, owing to the many different factors (both climatological and nonclimatological) that influence river hydrology and flood conditions. Research has also demonstrated a robust link between increased “wildfire weather” caused by climate change and more severe wildfires (e.g., increases in fuel aridity and acres burned) as a result of the strong effect of hotter, drier conditions on wildfire behavior.¹⁹⁷

Food and water security: Climate change has impaired food and water security in many regions of the world as a result of both slow-onset phenomena and extreme events.¹⁹⁸ Some of the slow-onset phenomena with the largest

192. *Id.*

193. *Id.*

194. *Id.* at 26.

195. *Id.* at 9.

196. IPCC AR6 WGI, *supra* note 6, at 120.

197. See, e.g., John T. Abatzoglou & A. Park Williams, *Impact of Anthropogenic Climate Change on Wildfire Across Western US Forests*, 113 PNAS 11770 (2016), <https://doi.org/10.1073/pnas.1607171113>.

198. IPCC AR6 WGII, *supra* note 181, at 9; ch. 4 (551–712).

impacts on food and water security include increasing temperatures, desertification, decreasing precipitation, land and forest degradation, and loss of biodiversity and ecosystem function. Extreme events such as acute droughts, floods, storms, and heat waves can also cause acute food and water stress (e.g., when drinking water systems are contaminated during storms). The effect of climate change on food and water security is evident across most regions of the world, particularly with respect to fisheries' yield and aquaculture production.¹⁹⁹

Public health and well-being: Attribution of public health outcomes from climate change can be challenging owing to data requirements and the complexity of isolating causal factors that contribute to health outcomes. As noted by Ebi et al. (2017), robust detection and attribution of health impacts require reliable long-term datasets, in-depth knowledge of the many drivers and confounding factors that affect public health outcomes, and refinement of analytic techniques to better capture the effect of anthropogenic forcing on health outcomes.²⁰⁰ Two key challenges are the fact that high-quality, long-term public health data are not available for many parts of the world and that there are many confounding factors that influence public health outcomes in any given region. Despite these limitations, the overall body of attribution research demonstrates a clear causal nexus between climate change and health outcomes.²⁰¹ IPCC AR6 expressed *very high confidence* in research showing that climate change is adversely affecting

199. *Id.* at 46, Fig. TS.3 (“Observed Global and Regional Impacts on Ecosystems and Human Systems Attributed to Climate Change”).

200. Kristie L. Ebi et al., *Detecting and Attributing Health Burdens to Climate Change*, 125 *Env't Health Persp.* 085004–1 (2017), <https://doi.org/10.1289/EHP1509>.

201. “[A]dvances are possible in the absence of complete data and statistical certainty: there is a place for well-informed judgments, based on understanding of underlying processes and matching of patterns of health, climate, and other determinants of human well-being.” *Id.* at 085004–7. To illustrate this point, the researchers discuss several contexts in which it is possible to show that a “proportion of the current burden of climate-sensitive health outcomes can be attributed to climate change”: (1) heat waves, (2) the emergence of tick vectors of Lyme disease in Canada, and (3) the emergence of *Vibrio* (bacteria) in northern Europe. For heat waves, the researchers described several approaches for estimating the number of heat wave deaths attributable to anthropogenic climate change. These included two variants on multistep attribution that would combine either the risk-based or storyline approach to extreme-event attribution with an assessment of how changes in exposure to heat waves affect mortality, as well as a single-step attribution approach, which would combine observations of the changes in the incidence and severity of heat waves with the exposure analysis. For *Vibrio*, the researchers found that it was possible to attribute increases in the incidence of *Vibrio* to incremental increases in sea surface temperatures, which could then be attributed to climate change. For tick vectors and Lyme disease, the researchers found that there was indirect evidence that higher temperatures were one of the forces leading to the expansion of these vectors, but that more detailed analyses of longer-term surveillance data were needed to actually quantify the relationship between climate change and tick vectors. One key takeaway from the authors of that study was that there are many different approaches to health impact attribution, including qualitative approaches. *See, generally, id.*

the physical and mental health of people around the world.²⁰² For example, extreme heat events are resulting in human mortality and morbidity across all regions (*very high confidence*),²⁰³ and studies on individual heat events have demonstrated that anthropogenic climate forcing can have a major effect on associated mortality and other health indicators.²⁰⁴ There are many other pathways through which climate change affects physical health. For example, climate change contributes to an increase in the prevalence of food-borne, water-borne, and vector-borne diseases that cause death and illness (*very high confidence to high confidence*).²⁰⁵ Climate change also exacerbates air pollution—for example, an increase in wildfire weather translates to increased wildfire smoke. Heat waves also contribute to air pollution, particularly in urban regions. The effects of climate change on air quality have been connected to increases in cardiovascular and respiratory disease.²⁰⁶

Cities, settlements, and infrastructure: The effects of climate change on human settlements and infrastructure are already apparent. IPCC AR6 found that sea-level rise, hydrological changes, permafrost thaw, and extreme events such as heat waves, droughts, wildfires, and floods are already causing disruptions of key infrastructure and services such as energy supply and transmission, communications, food and water supply (see above), and transportation systems (*high confidence*).²⁰⁷ In addition, cities and settlements are experiencing effects that have “extended from direct, climate-driven impacts to compound, cascading and systematic impacts (*high confidence*).” The nature and magnitude of these impacts vary considerably depending on the region, level of exposure to climate impacts, and vulnerability of affected populations and physical infrastructure. Coastal settlements and infrastructure are particularly vulnerable to compounding climate change impacts because of their exposure to sea-level rise, more intense storms and storm surge, and the associated risks of flooding, inundation, saltwater

202. IPCC AR6 WGII, *supra* note 181, at 11.

203. *Id.*

204. See, e.g., Daniel Oudin Åström et al., *Attributing Mortality from Extreme Temperatures to Climate Change in Stockholm, Sweden*, 3 *Nature Climate Change* 1050, 1051 (2013), <https://doi.org/10.1038/nclimate2022> (climate change doubled mortality during heat extremes in Sweden from 1980 to 2009); Daniel Mitchell et al., *Attributing Human Mortality During Extreme Heat Waves to Anthropogenic Climate Change*, 11 *Env’t Rsch. Letters* 1 (2016), <https://doi.org/10.1088/1748-9326/11/7/074006> (climate change had increased the risk of heat-related deaths during the 2003 European heat wave by approximately 70% in central Paris and 20% in London, and approximately 506 (± 51) deaths were attributable to climate change in Paris, and 64 (± 3) deaths were attributable in London).

205. IPCC AR6 WGII, *supra* note 181, at 11. See also *id.* at 50 (finding that changes in temperature, precipitation, and water-related disasters are linked to increased incidences of water-borne diseases such as cholera, especially in regions with limited access to safe water, sanitation, and hygiene infrastructure (*high confidence*)).

206. *Id.* at 11.

207. *Id.* at 53.

intrusion, and land subsidence.²⁰⁸ Many of these communities also depend on coastal and marine ecosystems for food security, livelihoods, and socioeconomic development—and they are thus uniquely affected by ecosystem alterations brought about by ocean warming, ocean acidification, and other aspects of climate change.²⁰⁹

Socioeconomic development and humanitarian impacts: All of the impacts described above have important implications for socioeconomic development, although attributing socioeconomic impacts to anthropogenic climate change can be tricky because of the fact that there are so many nonclimate factors that also influence development trajectories. Nonetheless, attribution studies have identified clear connections between both mean and extreme changes in the global climate system and adverse economic consequences across many regions and sectors, particularly in “climate-exposed” sectors and regions that experience high exposure and vulnerability to climate change impacts (*medium to high confidence*, depending on the sector and region).²¹⁰ Economic impacts are not uniform; some areas are disproportionately affected by adverse impacts,²¹¹ and some positive economic effects have been identified—for example, in “regions that have benefitted from lower energy demand as well as comparative advantages in agricultural markets and tourism” (*high confidence*).²¹² In addition, climate change is contributing to humanitarian disasters and human displacement, particularly in areas of high vulnerability to climate impacts (*high confidence*).²¹³ Extreme events are increasingly driving displacement across all regions of the world (*high confidence*), with disproportionate effects on Small Island Developing States

208. *Id.* (“Coastal cities are disproportionately affected by interacting, cascading and climate-compounding climate- and ocean-driven impacts, in part because of the exposure of multiple assets, economic activities and large populations concentrated in narrow coastal zones (*high confidence*)”).

209. See section titled “Socioeconomic development and humanitarian impacts.”

210. See IPCC AR6 WGII, *supra* note 181, at 11 (“Overall adverse economic impacts attributable to climate change, including slow-onset and extreme weather events, have been increasingly identified (*medium confidence*). . . . Economic damages from climate change have been detected in climate-exposed sectors, with regional effects to agriculture, forestry, fishery, energy, and tourism (*high confidence*), and through outdoor labour productivity (*high confidence*). Some extreme weather events, such as tropical cyclones, have reduced economic growth in the short-term (*high confidence*)”). See also *id.* at 54 (“The effects of climate change impacts have been observed across economic sectors, although the magnitude of the damage varies by sector and by region (*high confidence*). Recent extreme weather and climate-induced events have been associated with large costs through damaged property, infrastructure and supply chain disruptions, although development patterns have driven much of these increases (*high confidence*). Adverse impacts on economic growth have been identified from extreme weather events (*high confidence*) with large effects in developing countries (*high confidence*).”).

211. See, e.g., IPCC AR6 WGII, *supra* note 181, at 54 (discussing disproportionate adverse impacts on Small Island Developing States).

212. *Id.* at 11.

213. *Id.*

(*high confidence*).²¹⁴ Impacts on food and water security, discussed above, are also adversely affecting humanitarian outcomes and socioeconomic development.

Source Attribution

The field of source attribution encompasses research aimed at identifying and attributing climate change to specific sources. A source could be a particular actor (e.g., a country or a company), a sector, or an activity. In climate litigation, this research is used to answer questions such as: (1) whether and to what extent a particular entity has contributed to emissions and climate change impacts, (2) whether a source category generates a sufficient quantity of emissions such that a regulatory threshold is triggered, and (3) whether a government agency has adequately disclosed emissions and climate impacts from its projects and activities. As noted above, source attribution has been, and remains, a distinct discipline from what is commonly labeled “detection and attribution” in the climate science community. However, these research streams have begun to converge in recent years, and there is now a growing body of research aimed at linking specific entities (e.g., countries and private actors) to observed changes, such as sea-level rise and extreme events.²¹⁵

Methods and Parameters

The IPCC does not have a working group that serves as a direct analog for source attribution research like other issues discussed in this reference guide. IPCC Working Group III (WGIII) does synthesize some research on GHG emissions sources (e.g., estimates of regional and sectoral GHG contributions); however, the WGIII reports do not include data on the specific contributions of particular entities to global emissions, as may be required to support a legal finding of responsibility for or obligation to address climate change. Thus, source attribution data in climate litigation are often derived from sources other than IPCC reports.

214. *Id.*

215. See, e.g., David J. Frame et al., *Emissions and Emergence: A New Index Comparing Relative Contributions to Climate Change With Relative Climatic Consequences*, 14 *Env’t Rsch. Letters* 084009 (2019), <https://doi.org/10.1088/1748-9326/ab27fc>; R. Licker et al., *Attributing Ocean Acidification to Major Carbon Producers*, 14 *Env’t Rsch. Letters* 124060 (2019), <https://doi.org/10.1088/1748-9326/ab5abc>; Friederike E.L. Otto et al., *Assigning Historical Responsibilities for Extreme Weather Events*, 7 *Nature Climate Change* 757 (2017), <https://doi.org/10.1038/nclimate3419>; B. Ekwurzel et al., *The Rise in Global Atmospheric CO₂, Surface Temperature, and Sea Level from Emissions Traced to Major Carbon Producers*, 144 *Climatic Change* 579 (2017), <https://doi.org/10.1007/s10584-017-1978-0>.

The data used in source attribution come from direct measurements of emissions, which can be performed in situ or remotely from satellites, as well as documentary evidence of emissions contained in corporate reports, government inventories, and other sources.²¹⁶ Where direct emissions data are lacking, scientists can use indirect methods, such as models, to estimate emissions from sources and activities. Indirect methods are particularly important for estimating emissions from land use changes and nonpoint sources such as agricultural operations.

Establishing causation in the source attribution context involves quantifying the emissions contribution of the source and ascertaining the proportional contribution of those emissions to: (1) concentrations of greenhouse gases and other forcings and (2) how those changes in concentrations ultimately impact, for example, sea-level rise, extreme weather events, and the resultant impacts on ecosystems and/or communities. As noted above, there are some recent studies linking specific sources to global climate change and extreme events. However, most of the existing research on source attribution focuses on quantifying emissions from sources and determining the proportional contribution to increases in atmospheric greenhouse gases.

There are several factors that must be considered when making this calculation. First, climate change is not a product of a single pollutant or polluting activity, and different GHGs and other forcing agents have different effects on climate in terms of magnitude, duration, location, and type of effect. Second, there are some gaps in knowledge pertaining to total emissions of certain forcing agents, including historical emissions, as well as a good deal of uncertainty about the extent and timing of historical land use changes and their impact on atmospheric concentrations of greenhouse gases.²¹⁷ Some of these land use changes, like deforestation, also impact climate in other ways—for example, by altering the amount of sunlight converted to heat at the surface.²¹⁸

Nonetheless, scientists have endeavored to calculate the relative contributions of emissions and land use change, and, within the category of emissions, of

216. Because such reports are prepared by humans, sometimes pursuant to a political or social agenda, they may contain biases or errors of a different type than those found in raw data or instrumental records.

217. For additional context on the state of knowledge on different GHG emission sources and their relative contribution to global climate change, see NASEM, *Greenhouse Gas Emissions Information For Decision Making: A Framework Going Forward* (2022), <https://doi.org/10.17226/26641>.

218. For example, land use decisions that change the amount of sunlight absorbed at the surface can have an important or negligible effect on climate, depending on factors such as the latitude at which the deforestation occurs and the reflective properties of the surface underneath the previously forested area. Another complicating factor is that climate change itself directly impacts the magnitude of sources and sinks for greenhouse gases (e.g., a warmer ocean is less able to take up carbon dioxide, and changes in vegetation with climate change could switch some natural systems from net sources to net sinks, and vice versa).

different pollutants. In climate change attribution studies, scientists can bolster emissions data with actual measurements of atmospheric greenhouse gases (such as those taken at Mauna Loa in Hawaii) to determine the overall effect of human activity on climate, with the aforementioned caveats. Climate models can also be used both to evaluate the proportional contribution to atmospheric forcing agents and to ascertain the effect of that contribution on other aspects of climate change.

Finally, it is important to recognize that source attribution involves questions that cut across different social and scientific disciplines. Certainly, there is a physical science component to source attribution, as the ultimate goal is to ascertain the physical contribution of the source to anthropogenic climate change. But there are also social and normative questions that come into play when attributing emissions (or sequestration) to a particular source, particularly when trying to assign “responsibility” for emissions. Consider the many different ways that emissions can be “divvied up” across different lines—by stage of economic development, global region, country, sector, company, consumer, etc. Even within these categories, there are different ways of assigning emissions responsibility. For example, when assessing national responsibility for climate change, one can look at emissions that are generated within the country (territorial emissions), emissions embodied in products consumed within the country (consumption-based emissions), or emissions from fossil fuels produced within the country (production-based emissions). Similarly, when assessing corporate responsibility for climate change, there are important questions about the relative responsibility of upstream entities (e.g., fossil fuel producers) and downstream entities (e.g., manufacturers and end users of carbon-intensive products and consumers of fossil fuels) in addition to the entities that directly generate emissions.²¹⁹

Granted, it is entirely possible to avoid such normative questions when publishing information about source attribution. For example, a study could simply provide a breakdown of emissions across different countries using different accounting approaches (territorial, consumption-based, and production-based) without reaching any conclusions about the responsibility of different actors or source categories. But in practice, when attribution science is used in the courtroom, the question of responsibility may be of paramount importance.

International climate negotiations have historically focused on using national responsibility as the basis for allocating emission reduction burdens. This focus is evident in the United Nations Framework Convention on Climate Change (UNFCCC), which places the responsibility for reporting on and reducing

219. These types of considerations also factor into assessments of how to allocate the global carbon budget among different actors. See section titled “Impact Detection and Attribution” above.

emissions on national governments;²²⁰ the “Brazilian Proposal,” which emerged from UNFCCC negotiations in the mid-1990s and holds that greenhouse gas emission reduction targets should be set according to each country’s historical contribution to climate change;²²¹ and the Paris Agreement, which relies on nationally determined contributions (NDCs) as the primary basis for mitigating emissions.²²² The UNFCCC reporting framework has historically focused on territorial emissions (i.e., emissions generated within the country) as the metric for gauging national responsibility. However, in recent years there has been a strong push in both international and domestic forums to account for consumption-based emissions (i.e., emissions embodied in products consumed within the country) and extraction-based emissions (i.e., emissions from fossil fuels produced within the country) in addition to territorial emissions when assessing national obligations.²²³

While most national emissions inventories currently focus on territorial emissions, researchers have found that it would be relatively easy for countries to produce extraction-based and consumption-based inventories utilizing readily available data.²²⁴ In other words, pursuing these alternative accounting methodologies would not be significantly more expensive or technically challenging than the territorial approach. These alternative accounting methodologies also provide valuable insights that are not captured in the territorial approach—for

220. United Nations Framework Convention on Climate Change, May 9, 1992, S. Treaty Doc. No. 102–38, 1771 U.N.T.S. 107.

221. Emilio L. La Rovere et al., *The Brazilian Proposal on Relative Responsibility for Global Warming*, in *Building on the Kyoto Protocol: Options for Protecting the Climate* (Kevin A. Baumert et al. eds., 2002), <https://perma.cc/3ZJB-SQW5>.

222. Paris Agreement to the United Nations Framework Convention on Climate Change, Dec. 12, 2015, T.I.A.S. No. 16–1104.

223. For more information on these emissions accounting techniques, see Peter Erickson & Michael Lazarus, *Stockholm Environment Institute, Accounting for Greenhouse Gas Emissions Associated with the Supply of Fossil Fuels* (2013), <https://perma.cc/32VR-WLFQ>.

224. Glen P. Peters, *From Production-Based to Consumption-Based National Emission Inventories*, 65 *Ecological Econ.* 13 (2008), <https://doi.org/10.1016/j.ecolecon.2007.10.014>; Steven J. Davis & Ken Caldeira, *Consumption-Based Accounting of CO₂ Emissions*, 107 *PNAS* 5687 (2010), <https://doi.org/10.1073/pnas.0906974107>; Manfred Lenzen et al., *Building EORA: A Global Multi-Region Input–Output Database at High Country and Sector Resolution*, 25 *Econ. Sys. Res.* 20 (2013), <https://doi.org/10.1080/09535314.2013.769938>; Stavros Afionis et al., *Consumption-Based Carbon Accounting: Does It Have a Future?*, 8 *Wires Climate Change* 1 (2017), <https://doi.org/10.1002/wcc.438>; G.P. Peters et al., *A Synthesis of Carbon in International Trade*, 9 *Biogeosciences* 3247 (2012), <https://doi.org/10.5194/bg-9-3247-2012>; Kirsten S. Wiebe & Norihiko Yamano, *Estimating CO₂ Emissions Embodied in Final Demand and Trade Using the OECD ICIO 2015: Methodology and Results*, 2016/05 *OECD Sci., Tech. & Indus. Working Papers* 1 (2016), <https://doi.org/10.1787/5jlrcm216xkl-en>; Steven J. Davis et al., *The Supply Chain of CO₂ Emissions*, 108 *PNAS* 18554 (2011), <https://doi.org/10.1073/pnas.1107409108>; Thomas M. Power & Donovan S. Power, *The Impact of Powder River Basin Coal Exports on Global Greenhouse Gas Emissions*, Energy Found. (2013), <https://perma.cc/RV4A-SCH3>.

example, the consumption-based approach accounts for “leakage” of GHG emissions to other countries via trade and helps countries understand the importance of developing policies aimed at reducing consumption of carbon-intensive products. Ultimately, though they may carry different legal weight, all three methodologies are useful in addressing the question of who is “responsible” for climate change.

In addition, researchers have compiled detailed historical inventories of emissions attributable to some corporate actors and industries. Corporate emissions inventories and disclosures also serve as a source of data for assessing corporate contributions to climate change.

Status of Research

National contributions

Since the 1990s, countries, including the United States, have been preparing national GHG inventories pursuant to guidelines articulated by the UNFCCC. These inventories encompass emissions from energy, industrial processes and product use, agriculture, land use practices, and waste. Emission estimates are based on territorial emissions. The inventories are publicly available and frequently used as a data source by researchers who analyze emission trends. They provide somewhat comprehensive coverage of recent emissions from countries, but they do not provide a complete picture of historical emissions before the 1990s and early 2000s (for some countries). There are also accounting inconsistencies in the inventories owing to the issuance of different standards and reporting requirements for countries with different capacities.²²⁵

Governmental agencies, scientific organizations, and researchers have helped to fill these gaps through further research on national and sectoral emissions. Two key sources of data for this research are (1) atmospheric emission measurements, and (2) information about production and activities that contribute to GHG emissions. As noted by the National Academies, “[t]he most accurate and precise estimates are available for emission-producing activities that have explicit economic value (e.g., energy production),” and “[n]ational CO₂ emissions from fossil fuel burning are typically well characterized by economic data on fossil fuel trade and production.”²²⁶ In comparison, there is a good deal of uncertainty about CO₂ emissions estimates from land use and land use change,

225. For a more detailed discussion of limitations in UNFCCC emission inventories, see NASEM, *supra* note 217.

226. *Id.* at 50.

as well as emissions of non-CO₂ GHGs from both industry and land use.²²⁷ For example, recent research indicates that methane emissions from fossil fuel infrastructure are 50% larger than inventory estimates, and there is additional uncertainty regarding methane emissions from wetlands, reservoirs, agriculture, waste, and other sources.²²⁸ Nitrous oxide (N₂O) emissions, primarily from agriculture, are another important source of anthropogenic climate forcing, and measurements of atmospheric N₂O suggest that these emissions are underestimated globally.²²⁹

There are different accounting approaches for estimating national emissions. For example, researchers have developed national estimates of consumption- and extraction-based emissions in addition to territorial emissions.²³⁰ With respect to cumulative contributions, there is more uncertainty about the attribution of consumption-based emissions to individual countries as the result of limited data about historical emissions embodied in trade. Work has also been done to quantify emissions at different scales and to evaluate emissions by different metrics (e.g., per capita emissions within a country).²³¹ Overall, there is a large body of emissions data that litigants and other actors can draw on—especially for U.S. CO₂ emissions, which are well documented by both government agencies and independent researchers.²³²

There is also a growing body of research aimed at linking national emissions contributions to specific aspects of climate change. These studies use many of the same general techniques and climate models that are used when studying the aggregate effects of anthropogenic climate forcings; however, because these studies deal with a smaller quantity of emissions (national rather than global emissions), it can be more difficult to ascertain the impact of those emissions with precision and certainty. Accordingly, attribution studies for individual countries

227. *Id.* As a point of reference, the uncertainty for global fossil fuel CO₂ emissions is thought to be less than ±8%, whereas the uncertainty for land use CO₂ emissions is estimated at roughly ±75%. At the national level, these uncertainty estimates vary depending on the quality of data systems.

228. *Id.* See also NASEM, *Improving Characterization of Anthropogenic Methane Emissions in the United States* (2018), <https://doi.org/10.17226/24987>.

229. NASEM, *supra* note 217, at 51. See also Efisio Solazzo et al., *Uncertainties in the Emissions Database for Global Atmospheric Research (EDGAR) Emission Inventory of Greenhouse Gases*, 21 *Atmospheric Chemistry & Physics* 5655 (2021), <https://doi.org/10.5194/acp-21-5655-2021>.

230. See, e.g., Davis et al., *supra* note 224; Glen P. Peters et al., *Growth in Emission Transfers Via International Trade from 1990 to 2008*, 108 *PNAS* 8903 (2011), <https://doi.org/10.1073/pnas.1006388108>; Daniel Moran et al., *The Carbon Loophole in Climate Policy: Quantifying the Embodied Carbon in Traded Products* (2018), <https://perma.cc/2VDL-CGT2>.

231. H. Damon Matthews, *Quantifying Historical Carbon and Climate Debts Among Nations*, 6 *Nature Climate Change* 60 (2016), <https://doi.org/10.1038/nclimate2774>.

232. See, e.g., Hannah Ritchie et al., *United States: CO₂ Country Profile*, OurWorldinData (2020), <https://perma.cc/7WRT-XKSV> (cumulative emissions are important owing to the long atmospheric lifetime of CO₂); Hannah Ritchie, *Where in the World Do People Emit the Most CO₂?*, OurWorldinData (2019), <https://perma.cc/282X-UNUL>.

(or other individual sources) may result in a relatively broad range of possible values (i.e., a large confidence interval) as compared with attribution studies that deal with climate change in the aggregate. It is also important that such studies include a robust discussion of uncertainty.

Many national attribution studies deal with contributions to global mean surface temperature (GMST) increases, as there is robust physical understanding of the relationship between anthropogenic forcings and GMST.²³³ For example, Skeie et al. (2017) used a climate model to link the territorial emissions between 1850 and 2012 from multiple countries to GMST change, taking into account historical emissions and focusing on the largest emitters.²³⁴ In a similar study, Jones et al. (2023) calculated changes in GMST attributable to national CO₂, methane (CH₄), and nitrous oxide (N₂O) emissions from 1850 to 2021.²³⁵

Researchers are also examining the causal relationship between national emissions and a range of other impacts, including regional/country-specific impacts, and changes in extreme events. Otto et al. (2017) was the first study to apply the nation-based emissions framework to individual extreme-event attribution, focusing on a heat wave in Argentina.²³⁶ The researchers used two

233. See Nathan P. Gillett et al., *Constraining the Ratio of Global Warming to Cumulative CO₂ Emissions Using CMIP5 Simulations*, 26 J. Climate 6844 (2013), <https://doi.org/10.1175/JCLI-D-12-00476.1>; Pierre Friedlingstein & Susan Solomon, *Contributions of Past and Present Human Generations to Committed Warming Caused by Carbon Dioxide*, 102 PNAS 10832 (2005), <https://doi.org/10.1073/pnas.0504755102>; Ting Wei et al., *Developed and Developing World Responsibilities for Historical Climate Change and CO₂ Mitigation*, 109 PNAS 12911 (2012), <https://doi.org/10.1073/pnas.1203282109>; Julia Pongratz & Ken Caldeira, *Attribution of Atmospheric CO₂ and Temperature Increases to Regions: Importance of Preindustrial Land Use Change*, 7 Env't Rsch. Letters 034001 (2012), <https://doi.org/10.1088/1748-9326/7/3/034001>; Sophie C. Lewis et al., *Assessing Contributions of Major Emitters' Paris-Era Decisions to Future Temperature Extremes*, 46 Geophysical Rsch. Letters 3936 (2019), <https://doi.org/10.1029/2018GL081608>; Niklas Höhne et al., *Contributions of Individual Countries' Emissions to Climate Change and Their Uncertainty*, 106 Climatic Change 359 (2011), <https://doi.org/10.1007/s10584-010-9930-6>; Bo Fu et al., *The Contributions of Individual Countries and Regions to the Global Radiative Forcing*, 118 PNAS e2018211118 (2021), <https://doi.org/10.1073/pnas.2018211118>.

234. Ragnhild B. Skeie et al., *Perspective Has a Strong Effect on the Calculation of Historical Contributions to Global Warming*, 12 Env't Rsch. Letters 024022 (2017), <https://doi.org/10.1088/1748-9326/aa5b0a>. Skeie et al. noted that these findings were very sensitive to the parameters of the study, including technical decisions such as the time frame for the analysis, as well as more normative decisions about the basis for attributing emissions (e.g., place of extraction vs. place of burning vs. place of final consumption) and about whether to look at per capita or total emissions. They also emphasized that, in nonlinear systems, the proportional contribution to emissions will differ from the proportional contribution to impacts. *Id.*

235. Matthew W. Jones et al., *National Contributions to Climate Change Due to Historical Emissions of Carbon Dioxide, Methane, and Nitrous Oxide Since 1850*, 10 Sci. Data 155 (2023), <https://doi.org/10.1038/s41597-023-02041-1>. Note that Jones et al. and Skeie et al., *supra* note 234, use different time frames for the emissions analysis (1850–2012 and 1850–2021), which is one factor contributing to the different findings in these studies.

236. Otto et al., *supra* note 215.

different statistical methodologies²³⁷ to assess historical responsibility.²³⁸ This research provides useful initial insights on possible measures of national responsibility, but the large uncertainty ranges reported in the study illustrate the difficulty of attributing specific extreme events or climate impacts to specific emitters with a high level of precision.

Corporate contributions

Much of the research on corporate contributions to climate change has focused on corporations involved in fossil fuel production, energy, and other emissions-intensive industries. In the foundational Carbon Majors study, Heede (2014) developed estimates of historical emissions for ninety producers of oil, gas, coal, and cement based on their production records.²³⁹ Corporate emissions estimates can also be obtained through voluntary disclosures through initiatives like the Climate Disclosure Project, as well as legally mandated disclosures, which are becoming increasingly common.²⁴⁰ The IPCC and government agencies also compile emissions data for specific sectors (energy, transport, buildings, industry, forestry, agriculture, and waste). Sectoral data are potentially relevant when assessing government legal obligations related to the regulation of GHG emissions, or for comparing corporate emissions to industry norms.

Researchers have also begun conducting attribution studies aimed at linking corporate emissions to specific climate impacts (as with national emissions), for example, using findings from the Carbon Majors study to assess the impacts of emission contributions to global temperature change, sea-level rise,²⁴¹ and ocean acidification.²⁴² As noted in the discussion above, such studies rely on

237. As explained by the authors, the first methodology, the distribution method, “fits a distribution to the raw model data for both ensembles and estimates the percentage change in the distribution characteristics for individual regions by applying the [national] contributions to GMST.” The second methodology, the gradient method, “differs in that the return time curve is used directly to calculate the gradient of the curve at the threshold of the event in both ensembles, and scale between the two gradients according to the [percentage contributions].” *Id.*

238. The study also examined EU responsibility, but we focus on U.S. findings for the purposes of this reference guide. *Id.*

239. Richard Heede, *Tracing Anthropogenic Carbon Dioxide and Methane Emissions to Fossil Fuel and Cement Producers, 1854–2010*, 122 *Climatic Change* 229 (2014), <https://doi.org/10.1007/s10584-013-0986-y>; Richard Heede, *Carbon Majors: Accounting for Carbon and Methane Emissions 1854–2010: Methods and Results Report*, Climate Mitigation Servs. (2014), <https://perma.cc/3ZKX-Z7VZ>; Paul Griffin et al., *The Carbon Majors Database: Methodology Report 2017*, CDP (2017), <https://perma.cc/L7AR-BA89>.

240. See, e.g., EPA, Greenhouse Gas Reporting Program, <https://perma.cc/BV3G-JYQ9>; Calif. Air Resource Bd., Mandatory Greenhouse Gas Emissions Reporting, <https://perma.cc/L5A6-YTN4>.

241. Ekwurzel et al., *supra* note 215.

242. Licker et al., *supra* note 215.

well-established attribution techniques and climate models, but there tends to be more uncertainty (often expressed as a broader confidence interval) when attributing impacts to specific emitters and sources.

Projections of Future Climate Change

Climate change projections provide insights on the magnitude and scope of changes and impacts that could occur under different emission trajectories and warming scenarios. It is generally understood that the effects of climate change will become increasingly severe and pervasive as GHGs continue to accumulate in the atmosphere. However, the relationship among emissions, changes in the global climate system, and corresponding impacts is not always linear—for example, there are potential tipping points, feedback cycles, and cascading impacts that could result in acceleration of certain trends such as sea-level rise. Scientists have developed sophisticated methods for predicting future changes and impacts based on anthropogenic forcing and levels of atmospheric warming. Importantly, these projections are not intended to convey the world *as it will be* but rather the world *as it could be* depending on future human activities.

Methods and Parameters

Like attribution research, projections of climate change are based on physical understanding, climate datasets, statistical methods, and modeling tools. In addition, the findings from detection and attribution research, discussed above, provide the foundation for predicting future trends, insofar as they characterize the changes that are already underway as a result of anthropogenic climate forcing. Climate change projections are typically used in climate litigation to establish the foreseeability of future climate harms and to assess legal obligations to control GHG emissions and/or adapt to the effects of climate change.

The projections in IPCC assessments and many other studies rely heavily on the CMIP model ensemble to simulate how changes in climate forcers, such as GHGs, may affect the global climate system. IPCC AR6 relies on CMIP Phase 6 (CMIP6), which has better representations of physical processes as well as higher resolution compared to climate models used in previous IPCC assessments.²⁴³ The climate projections in IPCC AR6 are based on five illustrative scenarios that reflect potential changes in GHG emissions and land uses as well as natural climate forcers. They include near-term (2021–2040), mid-term (2041–2060), and long-term (2081–2100) projections.

243. IPCC AR6 WGI, *supra* note 6, at 12.

There is considerable confidence in the ability of global climate models to provide credible quantitative estimates of future climate change at large geographic scales.²⁴⁴ Confidence in global projections is higher for some climate variables (e.g., global mean temperature) than for others (e.g., precipitation). To evaluate the credibility and robustness of model projections, scientists will assess how well the model is able to simulate observed climate conditions. For example, the CMIP6 multimodel ensemble underwent a diagnostic evaluation that revealed that it captured most aspects of observed climate change well.²⁴⁵ One key improvement in global climate models is that they are now better at simulating natural climate variability, which can play a significant role in shaping future climate conditions.

Global climate models deal with large-scale weather patterns; they do not have the spatial resolution to simulate the effects of anthropogenic climate change at a regional or local scale. Scientists can use regional models to simulate the interactions between changes in large-scale weather patterns and regional or local conditions. Regional climate models have much finer spatial resolution (e.g., 1–50 km grid spacing, as opposed to 100–300 km grid spacing in a global model).²⁴⁶ There is generally more uncertainty and less confidence in projections at the regional scale because of the challenges with regional climate analysis and downscaling described above. In particular, internal variability tends to play a larger role in shaping regional climate conditions, and this complicates the identification of forced climate signals at the regional level. There are methods for reducing uncertainty in regional projections—for example, by combining and comparing results from different regional models, or by using initial-condition ensembles, where the same model is run repeatedly under identical forcing but with small variations in initial conditions (which helps researchers quantify uncertainty owing to internal variability).²⁴⁷

Even where models are unable to provide robust quantitative projections of future climate conditions, it may be possible to qualitatively predict trends with a fairly high level of confidence. Such qualitative projections can be used to assess climate risk and to define legal obligations related to climate change mitigation and adaptation. However, there are some variables for which it may be difficult to even reach qualitative conclusions about the direction of future trends. For example, it is unclear whether average precipitation will increase or decrease in some regions as a result of anthropogenic climate change.²⁴⁸

244. See *id.* at ch. 4. See also Randall et al., *supra* note 48.

245. IPCC AR6 WGI, *supra* note 6, at 561.

246. Eric P. Salathé Jr. et al., *Regional Climate Model Projections for the State of Washington*, 102 *Climatic Change* 51 (2010), <https://doi.org/10.1007/s10584-010-9849-y>.

247. IPCC AR6 WGI, *supra* note 6, at 562.

248. At the global scale, models project that there will be greater contrast in mean precipitation between dry and wet regions (i.e., dry areas will become drier; wet areas will become wetter) but with large regional variations and low confidence in projections. *Id.* at 1109. Some regions, such

As noted above, climate projections must be issued in reference to future emissions (or anthropogenic forcings) and/or warming scenarios. These future emissions, which will depend primarily on future human decisions, thus include an additional type of uncertainty relative to attribution research. In its latest assessment, the IPCC used the following emission scenarios based on Shared Socio-economic Pathways (SSPs) to frame its analysis: (1) a “very low” emissions scenario, where GHG emissions decline sharply, reaching net zero around 2050, followed by net negative emissions (SSP1–1.9); (2) a “low” emissions scenario, where emissions decline quickly (but not as fast as the very low scenario), reaching net zero emissions after 2050, followed by net negative emissions (SSP1–2.6); (3) an “intermediate” emissions scenario, where CO₂ emissions remain at current levels until mid-century and then decline to net zero by 2100 (SSP2–4.5); (4) a “high” scenario where emissions continue to increase through the century (SSP3–7.0); and (5) a “very high” scenario with larger emission increases, followed by a slight decline in emissions toward the end of the century (SSP5–8.5).²⁴⁹ These five scenarios correspond with different temperature increases (see Figure 2), and they have been used as the basis for climate modeling in CMIP6 ensembles.

Status of Research

Using the emissions scenarios described above, IPCC AR6 projected future changes in global surface temperature²⁵⁰ and other aspects of the climate system, including (1) mean sea level, (2) ocean acidification and deoxygenation, (3) intensity and variability of the global hydrological cycle, (4) the extent and volume of all cryosphere resources such as sea ice and glaciers, and (5) climatological extremes.²⁵¹ It also projected changes in climate hazards and impacts under future emissions scenarios, such as changes in heat-related mortality and morbidity; food-borne, water-borne, and vector-borne diseases; flooding in coastal and low-lying areas; biodiversity loss; impacts on food and water security; and mental health impacts.²⁵² IPCC AR6 concluded that the loss and damage caused by anthropogenic climate change will be severe even if we limit global warming to 1.5°C or “well below” 2.0°C (the Paris Agreement targets for averting dangerous anthropogenic climate change, which are often used to formulate

as the western United States, are projected to have increased precipitation in some models and decreased precipitation in other models. *Id.* at 1110 (Fig. 8.14). However, this is not the case for most regions. There are other regions for which there is more pervasive uncertainty about the trajectory of future impacts (e.g., ice storms and tornadoes).

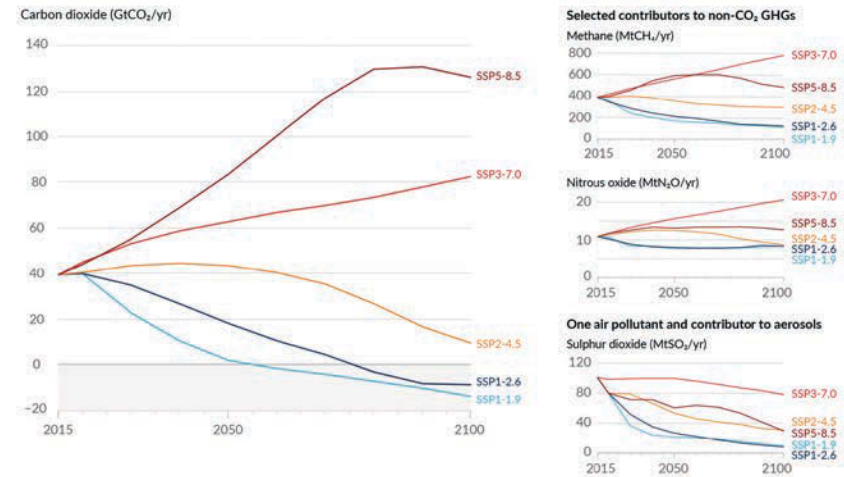
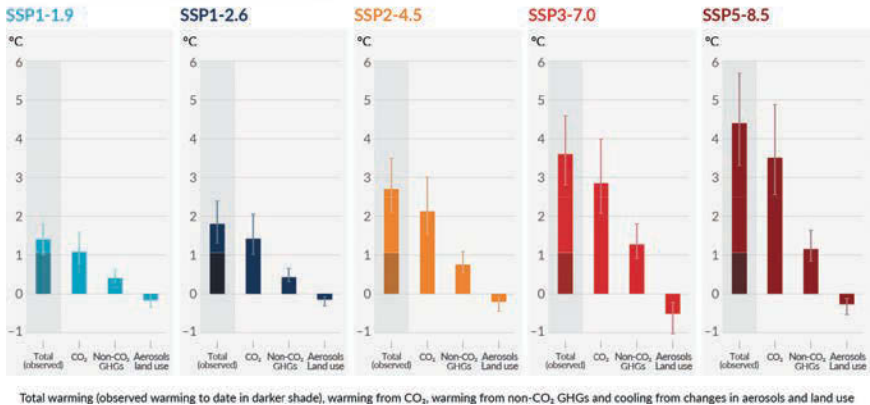
249. IPCC AR6 WGI, *supra* note 6, at 232 (Cross-Chapter Box 1.4).

250. *Id.* at 13–14.

251. *Id.* See also IPCC AR6 Synthesis Report, *supra* note 150.

252. *Id.*

Figure 2. Emission scenarios in IPCC AR6.

(a) Future annual emissions of CO₂ (left) and of a subset of key non-CO₂ drivers (right), across five illustrative scenarios(b) Contribution to global surface temperature increase from different emissions, with a dominant role of CO₂ emissions
Change in global surface temperature in 2081–2100 relative to 1850–1900 (°C)

Source: Figure SPM.4 in IPCC, 2021: Summary for Policymakers. In: *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* [Masson-Delmotte, V., P. Zhai, A. Pirani, S.L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M.I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J.B.R. Matthews, T.K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou (eds.). Cambridge University Press, Cambridge, UK, and New York, NY, USA, pp. 3–32, doi: 10.1017/9781009157896.001.]

scenarios for climate projections and shape emissions reduction efforts).²⁵³ Many of the projected changes in the climate system will be “irreversible on centennial to millennial timescales,” particularly changes in the ocean, ice sheets, and global sea level.²⁵⁴

Notably, for any given future warming level, many climate-related risks are higher than those assessed in AR5, and the severity of projected long-term impacts is often significantly higher than currently observed impacts like those described above (*high confidence*).²⁵⁵ The effects of climate change will also interact with nonclimatic risks, creating “compound and cascading risks that are more complex and difficult to manage” (*high confidence*).²⁵⁶ The relationship between impacts and global warming is often not linear, and small temperature increases can result in dramatically more severe impacts. IPCC AR6 also found that the planet is already very close to surpassing many critical thresholds within the climate system, and some of these thresholds will likely be surpassed if global warming exceeds 1.5°C.²⁵⁷ More recent studies on tipping points that were published after evidence was compiled for IPCC AR6 indicate that we are close to surpassing certain thresholds, and may have already surpassed some, particularly those associated with cryosphere impacts (e.g., the melting of ice sheets).²⁵⁸

Adaptation measures can play a significant role in mitigating certain risks, such as the risks associated with extreme precipitation and flooding. However, adaptation may not be as effective at mitigating other harmful impacts, such as those on biodiversity and ecosystems. Moreover, the effect of climate change on vulnerable populations and ecosystems often reduces their adaptive capacity, thus creating a compounding problem where adaptation becomes increasingly challenging and costly as climate change becomes more severe. Additionally, most adaptations involve tradeoffs, and there is a risk of maladaptation and inequitable adaptation.

How Climate Science Factors into Litigation

The United States has seen considerable growth in the field of climate litigation, that is, cases that involve legal claims related to climate change mitigation,

253. IPCC, *Global Warming of 1.5°C: An IPCC Special Report on the Impacts of Global Warming of 1.5°C Above Pre-Industrial Levels and Related Global Greenhouse Gas Emission Pathways 5* (Valérie Masson-Delmotte et al. eds., 2018), <http://doi.org/10.1017/9781009157940.001> [hereinafter IPCC 1.5°C Report].

254. IPCC AR6 Synthesis Report, *supra* note 150, at 24.

255. *Id.* at 14.

256. *Id.*

257. *Id.* at 77.

258. See, e.g., McKay et al., *supra* note 34. Once a tipping point is reached, it can take hundreds or thousands of years to reach a new equilibrium.

adaptation, compensation, and disclosures.²⁵⁹ This section provides an overview of the types of climate cases that are being filed in U.S. courts, particularly federal courts, and describes how different areas of climate science may factor into judicial assessments of causation, foreseeability, and legal obligations and authorities. It also discusses some considerations that are relevant when evaluating the admissibility, probative value, and weight of scientific evidence and expert testimony on matters related to climate change.

Overview of Climate Litigation

Many different types of climate-related claims have been filed in federal and state courts over the past several decades. Most of the federal cases involve questions about government obligations or authorities with respect to climate change mitigation, adaptation, and disclosures, as detailed below. These are typically administrative law claims, where judges may encounter questions about climate science when assessing (1) the scope of scientific evidence that must be considered by the agency before it decides on a course of action and (2) whether the agency has made a decision based on that evidence that meets the applicable legal standard.²⁶⁰ There are some exceptions, including constitutional claims against governments²⁶¹ and statutory and tort law claims against corporate defendants.²⁶² But the majority of federal claims involve government defendants and tend to fall in one of the following broad categories.

259. Mitigation refers to actions to reduce greenhouse gas emissions and other anthropogenic drivers of climate change. Adaptation refers to actions to prepare for and respond to the effects of climate change on human and natural systems. Compensation refers to payments or other resources given to people who have experienced damages as a result of anthropogenic climate change. Disclosures refer to the analysis and disclosure of information related to climate change (particularly climate-change-related risks) in government and corporate documents, such as environmental impact statements and financial reports.

260. *See, e.g.*, 5 U.S.C. § 706(2)(A) (directing courts to hold unlawful and set aside agency action, findings, and conclusions that are “arbitrary, capricious, an abuse of discretion, or otherwise not in accordance with law”); 5 U.S.C. § 706(2)(E) (directing courts to set aside a formal rulemaking or adjudication if it is not supported by “substantial evidence”). *See also* *Motor Vehicles Mfrs. Ass’n v. State Farm Mut. Auto. Ins. Co.*, 463 U.S. 29, 43 (1983) (when applying the arbitrary and capricious standard, courts evaluate whether the agency “examine[d] the relevant data and articulate[d] a satisfactory explanation for its action, including a rational connection between the facts found and the choice made”).

261. *See infra* note 273.

262. For example, federal cases have been filed against corporate actors seeking to establish liability for climate-change-related damages. *See, e.g.*, *Native Village of Kivalina v. ExxonMobil Corp.*, 696 F.3d 849 (9th Cir. 2012); *Complaint, Municipalities of Puerto Rico v. ExxonMobil Corp.*, No. 3:22-cv-01550 (D.P.R. Nov. 22, 2022). However, most corporate liability cases in the United States involve state law claims. *See, e.g.*, *City of Oakland v. BP PLC*, No. 22-16810, 2023

First, there are legal disputes about the scope of government obligations or authorities to regulate GHG emissions (i.e., mitigation). This category includes administrative lawsuits seeking regulatory action for GHG emissions pursuant to the Clean Air Act or other federal and state statutes, as well as lawsuits where agencies are defending GHG regulations against industry and/or state challenges.²⁶³ It also includes constitutional claims alleging that the government has violated fundamental rights through inadequate control of GHG emissions (these are sometimes called “atmospheric trust” cases, as the public trust doctrine is typically invoked alongside substantive due process).²⁶⁴ Constitutional claims have also been filed in opposition to government actions aimed at controlling GHG emissions and activities that contribute to GHG emissions (e.g., government restrictions on fossil fuel infrastructure).²⁶⁵

Second, many of the federal lawsuits filed against government agencies involve questions about procedural obligations to analyze and disclose climate-change-related considerations in administrative decision-making. These typically take the form of Administrative Procedure Act and National Environmental Policy Act (NEPA) claims alleging that federal agencies have not adequately accounted for climate-change-related considerations in administrative documentation prior to undertaking final agency action.²⁶⁶ Cases in this category may focus on agency obligations to account for GHG emissions and associated impacts attributable to federal actions, such as fossil fuel leases, natural gas pipeline approvals,

WL 8179286 (9th Cir. Nov. 27, 2023) (affirming remand to state court); *City & Cnty. of Honolulu v. Sunoco LP*, 39 F.4th 1101 (9th Cir. 2022), *cert. denied*, 143 S. Ct. 1795 (2023) (affirming remand to state court). Corporate actors may also be sued for the “failure to adapt” to climate change (e.g., the failure to prepare a coastal facility that stores hazardous substances for the effects of sea-level rise and more severe storms), or the “failure to disclose” known climate risks (e.g., in federal securities filings). *See, e.g.*, *Complaint, Conservation Law Found. v. Shell Oil Prods. US*, No. 1:17-cv-00396 (D.R.I. Aug. 28, 2017); *Complaint, Conservation Law Found. v. ExxonMobil Corp.*, No. 1:16-cv-11950 (D. Mass. Sept. 29, 2016).

263. *See, e.g.*, *Massachusetts v. EPA*, 549 U.S. 497 (2007); *Coal. for Responsible Regul. v. EPA*, 684 F.3d 102 (D.C. Cir. 2012), *aff’d in part, rev’d in part on other grounds by Util. Air. Regul. Grp. v. EPA*, 573 U.S. 302 (2014).

264. *See, e.g.*, *Juliana v. United States*, No. 6:15-cv-01517, 2023 WL 9023339 (D. Or. Dec. 29, 2023); *Animal Legal Defense Fund v. United States*, No. 19–35708, 2022 WL 5241274 (9th Cir. Feb. 9, 2022); *Komor v. United States*, No. 22–15851, 2023 WL 4313136 (9th Cir. July 3, 2023). State courts have also confronted constitutional and public trust claims about government obligations with regard to GHG emissions. *See, e.g.*, *Held v. Montana*, No. CDV-2020-307, 2023 WL 5229257 (Mont. Dist. Ct. Aug. 14, 2023).

265. *See, e.g.*, *Montana v. City of Portland*, No. 3:23-cv-00219, 2023 WL 8452447 (D. Or. Oct. 12, 2023); *Levin Richmond Terminal Corp. v. City of Richmond*, 482 F. Supp. 3d 944 (N.D. Cal. 2020).

266. *See, e.g.*, *Diné Citizens Against Ruining Our Env’t v. Haaland*, 59 F.4th 1016 (10th Cir. 2023); *Ctr. for Biological Diversity v. Nat’l Highway Traffic Safety Admin.*, 538 F.3d 1172, 1189 (9th Cir. 2008).

and forest management decisions.²⁶⁷ They may also focus on agency obligations to account for the effects of climate change on the action and its affected environment—for example, parties to a lawsuit may argue that an agency is obliged to account for observed and predicted hydrological changes in its assessment of a water management project.²⁶⁸

Third, there are cases that deal with government obligations or authorities to prepare for and respond to the effects of climate change on physical infrastructure, natural resources, or other matters under agency jurisdiction (i.e., adaptation). For example, there have been over 100 Endangered Species Act (ESA) cases that involve climate-change-related claims.²⁶⁹ These include cases where plaintiffs are trying to compel the federal government to conduct assessments of climate-change-related risks and/or issue protections for climate-change-imperiled species, as well as cases where the government is defending ESA protections for climate-change-imperiled species against legal challenges.²⁷⁰ Government agencies may also be sued regarding the treatment of climate change and climate science in decisions about land management plans, quotas for the sustainable use of resources, and government approvals for infrastructure and facilities that may be affected by climate change.²⁷¹ Finally, there are some examples of lawsuits aimed at blocking government adaptation measures (e.g., on the grounds that such measures constitute a regulatory taking or are in violation of federal law).²⁷²

267. See, e.g., *Diné Citizens*, 59 F.4th at 1016; 350 *Montana v. Haaland*, 50 F.4th 1254 (9th Cir. 2022); *Sierra Club v. Fed. Energy Regul. Comm'n*, 867 F.3d 1357 (D.C. Cir. 2017); *Ctr. for Biological Diversity v. U.S. Forest Serv.*, No. CV 22-114-M-DWM, 687 F. Supp. 3d 1053 (D. Mont. 2023); *High Country Conservation Advocs. v. U.S. Forest Serv.*, 333 F. Supp. 3d 1107 (D. Colo. 2018). Some cases specifically challenge the failure to employ a “social cost of carbon” metric to estimate the significance of GHG emissions. See, e.g., *Ctr. for Biological Diversity v. Fed. Energy Regul. Comm'n*, 67 F.4th 1176 (D.C. Cir. 2023); *Utah Physicians for a Healthy Env't v. U.S. Bureau of Land Mgmt.*, 528 F. Supp. 3d 1222, 1228 (D. Utah 2021).

268. See, e.g., *AquAlliance v. U.S. Bureau of Reclamation*, 287 F. Supp. 3d 969 (E.D. Cal. 2018).

269. Sabin Ctr. for Climate Change Law, *Federal Statutory Claims: Endangered Species Act and Other Wildlife Protection Statutes*, U.S. Climate Change Litig., <https://perma.cc/B9RT-VH3V>.

270. See, e.g., *Alaska Oil & Gas Ass'n v. Ross*, 722 F. App'x 666 (9th Cir. 2018); *Ctr. for Biological Diversity v. Zinke*, 900 F.3d 1053 (9th Cir. 2018); *Alaska Oil & Gas Ass'n v. Pritzker*, 840 F.3d 671 (9th Cir. 2016); *In re Polar Bear Endangered Species Act Listing & 4(d) Rule Litig.*, 709 F.3d 1 (D.C. Cir. 2013); *Buffalo Field Campaign v. Williams*, 579 F. Supp. 3d 186 (D.D.C. 2022), *appeal dismissed sub nom. Buffalo Field Campaign v. Haaland*, No. 22-5064, 2022 WL 2135456 (D.C. Cir. June 14, 2022)).

271. See, e.g., *Maine Lobsterman's Ass'n, Inc. v. Nat'l Marine Fisheries Serv.*, 70 F.4th 582 (D.C. Cir. 2023); *New York v. Raimondo*, 594 F. Supp. 3d 588 (S.D.N.Y. 2022); *Sound Action v. U.S. Army Corps of Eng'rs*, No. C18-0733JLR, 2019 WL 446614 (W.D. Wash. Feb. 5, 2019); *Ctr. for Biological Diversity v. Nat'l Marine Fisheries Serv.*, 977 F. Supp. 2d 55 (D.P.R. 2013), *as amended* (Oct. 23, 2013), *adhered to on reconsideration*, 191 F. Supp. 3d 157 (D.P.R. 2016).

272. See, e.g., *Complaint, Casa Mira Homeowners Ass'n v. Cal. Coastal Comm'n*, No. 19-civ-04677 (Cal. Super. Ct. Dec. 13, 2019); *Complaint, Jurisich Oysters, LLC v. U.S. Army Corps of Eng'rs*, No. 2:24-cv-00106 (E.D. La. Jan. 11, 2024).

These three categories encompass most but not all of the federal climate claims that have been filed to date. Some examples of lawsuits falling outside of these categories include: (1) legal challenges to federal and state regulations requiring disclosure of GHG emissions and climate-related risks from private companies;²⁷³ (2) legal challenges to the federal government’s use of “social cost of carbon” metrics in regulatory programs and other agency actions;²⁷⁴ (3) legal challenges to regulatory approvals of clean energy projects;²⁷⁵ (4) lawsuits seeking to establish liability on the part of corporations for climate-change-related damages;²⁷⁶ and (5) lawsuits seeking to establish obligations on the part of corporate actors to prepare for and respond to the effects of climate change on facilities, operations, or resources under their control.²⁷⁷

Legal Applications of Climate Science

There are several ways in which climate science may factor into the resolution of climate lawsuits. First, climate science may be used to evaluate claims of causation and harm, e.g., whether the GHG emissions at issue in a case cause or contribute to public endangerment or injury to a specific plaintiff. Second, climate science may be used to evaluate whether climate-related risks or harms were or are foreseeable. Finally, climate science may be used to determine the scope of a defendant’s legal obligations and authorities with regard to GHG mitigation, climate change adaptation, climate-related disclosures, and other matters.

Causation and Harm

Disputes about causation and harm arise in different types of climate lawsuits. For example, in cases where plaintiffs are seeking to compel government regulation of GHG emissions from a sector or activity, the parties may dispute the nature and/or magnitude of harm that can be attributed to those emissions, and whether the harm is substantial enough to trigger a regulatory obligation on the part of the government defendant.²⁷⁸ Causation may also be an issue in cases

273. See, e.g., *Iowa v. Sec. & Exch. Comm’n*, No. 24–1522 and consolidated cases (8th Cir. Mar. 12, 2024); *Chamber of Commerce v. California*, No. 2:24-cv-00801 (C.D. Cal. Jan. 30, 2024).

274. See *Missouri v. Biden*, 52 F.4th 362 (8th Cir. 2022), *cert. denied*, 144 S. Ct. 278 (2023).

275. See, e.g., *Seafreeze Shoreside, Inc. v. U.S. Dep’t of the Interior*, No. 1:22-CV-11091-IT, 2023 WL 6691015 (D. Mass. Oct. 12, 2023).

276. See *supra* note 262.

277. *Id.*

278. See, e.g., *Juliana v. United States*, 339 F. Supp. 3d 1062, 1096 (D. Or. 2018), *rev’d and remanded*, 947 F.3d 1159 (9th Cir. 2020). In the foregoing case, a group of youth plaintiffs claimed

where government agencies are defending GHG regulations against legal challenges, insofar as it may be necessary to demonstrate that those emissions cause some form of environmental or public harm in order to justify the regulatory action.²⁷⁹ Finally, the issue of causation and harm tends to be a major element of the standing analysis in cases where the plaintiff is alleging that they have experienced an injury as a result of the defendant's actions or inaction with regard to GHG emissions and climate change.²⁸⁰

As discussed in the sections of this reference guide titled “Foundational Components of Climate Science” and “Climate Change Detection, Attribution, and Projections,” there is scientific agreement that human GHG emissions are the dominant cause of global climate change. Thus, causation disputes in recent climate cases typically deal with one or both ends of the causation chain linking emissions to impacts (i.e., questions of source attribution and impact attribution). The source attribution questions pertain to the defendant's contribution to climate change—for example, what are the emissions attributable to the defendant's conduct, and do those emissions surpass a threshold of materiality or substantiality as may be necessary to establish legal causation? The impact attribution questions pertain to the nature of the alleged harm—for example, has the plaintiff experienced a specific and concrete injury that is attributable to climate change?²⁸¹ These two sets of questions combine into the overarching question of whether the defendant is responsible for the climate-change-related harms that are at issue in the case.

The contours of this analysis will differ depending on the nature of the claim and the status of the litigants. One important consideration is whether the case requires a showing of particularized injury as opposed to generalized

that the federal government had violated their constitutional rights by permitting, authorizing, and subsidizing fossil fuel extraction and consumption that contributed to climate change. They sought both declaratory and injunctive relief requiring the federal government to reduce emissions from fossil fuels. The district court held that plaintiffs had Article III standing because they had alleged sufficiently personalized and concrete injuries that were fairly traceable to the GHG emissions resulting from U.S. fossil fuel production and use. The Ninth Circuit reversed, holding that plaintiffs had failed to satisfy the redressability prong of standing because the injunctive relief they sought was outside the power of an Article III court.

279. See, e.g., *Coalition for Responsible Regulation v. EPA*, 684 F.3d 102 (D.C. Cir. 2012).

280. See, e.g., *Massachusetts v. EPA*, 549 U.S. 497 (2007); *Wash. Env't Council v. Bellon*, 732 F.3d 1131 (9th Cir. 2013), *reh'g en banc denied*, 741 F.3d 1075 (9th Cir. 2014); *Comer v. Murphy Oil USA*, 585 F.3d 855 (5th Cir. 2009), *rev'd and remanded in part, on other grounds*, 585 F.3d 855 (5th Cir. 2009); *Amigos Bravos v. U.S. Bureau of Land Mgmt.*, 816 F. Supp. 2d 1118, 1136 (D.N.M. 2011); *Juliana*, 339 F. Supp. 3d at 1062.

281. These questions of source attribution may implicate questions about climate change and extreme-event attribution. For example, if a plaintiff alleges injury based on exposure to an extreme event, the court would need to refer to extreme-event attribution research to evaluate the relationship between global climate change and the extreme event (what is the attributable risk?), as well as impact attribution data to evaluate the relationship between the plaintiff's injury and the event.

harm. A showing of particularized injury is required for standing. Specifically, for Article III standing, plaintiffs must demonstrate that: (1) they have experienced an injury-in-fact that is concrete and particularized, and actual or imminent, not conjectural or hypothetical; (2) the injury-in-fact is fairly traceable to the defendants' allegedly unlawful actions; and (3) the injury could be redressed by a favorable court decision.²⁸² Plaintiffs who are suing on the basis of climate-change-related injuries may attempt to establish these elements by presenting evidence that they are experiencing a unique injury as a result of climate change and that the defendant has made a substantial enough contribution to global climate change such that the injury can be fairly traced to the defendant's conduct.²⁸³

The Supreme Court's 5–4 decision in *Massachusetts v. EPA* is illustrative.²⁸⁴ The state of Massachusetts and other petitioners had filed a lawsuit challenging EPA's denial of a petition to regulate GHG emissions from motor vehicles under the Clean Air Act. One of EPA's stated rationales for declining the petition was that "a causal link between greenhouse gases and the increase in global surface air temperatures was not unequivocally established."²⁸⁵ EPA also argued that the petitioners lacked standing because they could not demonstrate a particularized injury that could be fairly traced to its decision to not regulate GHG emissions from motor vehicles. However, the Supreme Court disagreed, finding that:

The harms associated with climate change are serious and well recognized. The Government's own objective assessment of the relevant science and a strong consensus among qualified experts indicate that global warming threatens, *inter alia*, a precipitate rise in sea levels, severe and irreversible changes to natural ecosystems, a significant reduction in winter snowpack with direct and important economic consequences, and increases in the spread of disease and the ferocity of weather events. That these changes are widely shared does not minimize Massachusetts' interest in the outcome of this litigation.²⁸⁶

282. *Lujan v. Defs. of Wildlife*, 504 U.S. 555 (1992). In addition to these constitutional requirements, there is also a prudential requirement that the complaint be in the zone of interests of the applicable statute in order for the plaintiff to have standing. *Bennett v. Spear*, 520 U.S. 154 (1997).

283. The second and third elements of standing (causation and redressability) are closely related, sometimes referred to as "two sides of the same coin." See *FDA v. All. for Hippocratic Med.*, 602 U.S. 367, 380–81 (2024); *Ctr. for Biological Diversity v. EPA*, 90 F. Supp. 3d 1177, 1190 (W.D. Wash. 2015); *Gonzales v. Gorsuch*, 688 F.2d 1263, 1267 (9th Cir. 1982); *Duke Power Co. v. Carolina Env't Study Grp.*, 438 U.S. 59, 74 (1978). Thus, if plaintiffs establish that their injury can be "fairly traced" to the defendant's conduct, then there is a reasonable prospect that a court order that compels a change in that conduct will at least partially mitigate the injury. However, if the court does not have the authority to issue a remedy, then it may find that plaintiffs lack standing based on this third element. See section titled "Legal Obligations and Authorities" below.

284. *Massachusetts v. EPA*, 549 U.S. 497 (2007).

285. *Id.* at 497.

286. *Id.* at 499.

The Court also went on to find that Massachusetts had suffered a particularized and cognizable injury based on documentary evidence and expert testimony describing climate-related impacts such as the loss of state-owned property to rising sea levels; the added costs to deal with emergency response measures caused by more frequent intense storm surge flooding events; damage to state-owned historic, archaeological, and natural resources; and damage to state-owned facilities and infrastructure along the coast.²⁸⁷ The Court held that this injury could be fairly traced to EPA's decision not to regulate based on IPCC reports, which demonstrated a clear causal nexus between GHG emissions and climate change impacts as a general matter, and emissions data showing that the unregulated source category (U.S. motor vehicles) generated approximately 1.7 billion metric tons of CO₂ in 1999 (more than 6% of worldwide emissions). The Court noted that this was a "meaningful contribution" to global emissions by any metric.²⁸⁸ With regard to redressability, the Court acknowledged that Clean Air Act regulation would not fully mitigate the problem, but the prospect of partial redress was all that was required for Article III standing.²⁸⁹

Turning to the merits of the lawsuit, the Court held that GHG emissions qualified as "air pollutants" within the meaning of the Clean Air Act and that EPA had an obligation to regulate those emissions unless it "determine[d] that greenhouse gases do not contribute to climate change" or otherwise provided "some reasonable explanation as to why it cannot or will not exercise its discretion to determine whether they do."²⁹⁰ Although the Court remanded to EPA to make its own determination on these issues, the Court's own decision contained ample information indicating that those GHG emissions did endanger public health and welfare (e.g., the Court's assertion that the "harms associated with climate change are serious and well recognized" and the Court's finding that GHG emissions from motor vehicles qualified as a meaningful contribution to global climate change). EPA issued an affirmative endangerment finding for GHG emissions from this source category shortly after the Court issued this judgment, which was subsequently challenged and upheld in *Coalition for Responsible Regulation v. EPA*, discussed below.²⁹¹

287. The Court noted that Massachusetts had a "special position and interest" in the case, in part because "it actually owns a great deal of the territory alleged to be affected" by climate change, and in part because of its status as a sovereign state. *Id.* at 520–23. The Court referred to data in the petitioners' affidavits showing that "global sea levels rose between 10 and 20 centimeters over the 20th century as a result of global warming and have already begun to swallow Massachusetts' coastal land" and that "[r]emediation costs alone . . . could reach hundreds of millions of dollars." *Id.* at 521–23.

288. *Id.* at 525.

289. *Id.* at 525–26.

290. *Id.* at 501 ("EPA can avoid promulgating regulations only if it determines that greenhouse gases do not contribute to climate change or if it provides some reasonable explanation as to why it cannot or will not exercise its discretion to determine whether they do.").

291. Shortly before publication of this Manual, EPA issued a proposed rule to rescind the endangerment finding and associated regulations for GHG emissions from motor vehicles. See Reconsideration of 2009 Endangerment Finding and Greenhouse Gas Vehicle Standards, 90 Fed.

Courts may confront questions about whether the GHG emissions at issue in a case are substantial enough to cause legally cognizable harm when adjudicating both standing and the merits of cases (i.e., do the emissions represent a “meaningful” or “material” contribution to climate-change-related harms?).²⁹² The threshold of legal relevance, significance, or materiality will depend on the nature of the claim and the applicable legal standard.²⁹³ Attribution research and predictive research can provide technical insights to help answer this question, including insights on the scope and magnitude of impacts that could be attributed to those emissions, but this is ultimately a legal determination.

Questions about whether specific harms can be attributable to global climate change may also arise as part of the standing or the merits analysis. The evidence required to demonstrate such causation depends on the type of injury and, for standing purposes, the scope of the plaintiff’s affected interests. Attribution data tend to be more robust for impacts at a broader spatial and temporal scale, and for certain types of impacts (e.g., heat-related impacts). A plaintiff like the state of Massachusetts will experience many climate impacts at a relatively broad spatial scale, making it easier to demonstrate that at least some of those injuries can be attributed to climate change with reasonable confidence. This is also true for many cities and communities—the breadth of local interests affected by climate change can be quite large, and some communities are directly exposed to trends and events that have been attributed to climate change with high confidence.²⁹⁴ It

Reg. 36288 (proposed Aug. 1, 2025) (to be codified at 40 C.F.R. pts. 85, 86, 600, 1036, 1037, 1039). EPA also issued a proposed rule to rescind the “cause or contribute” findings and associated regulations for GHG emissions from fossil fuel-fired power plants. See Repeal of Greenhouse Gas Emissions Standards for Fossil Fuel-Fired Electric Generating Units, 90 Fed. Reg. 25752 (proposed June 17, 2025) (to be codified at 40 C.F.R. pt. 60). EPA initiated both rulemaking processes pursuant to Exec. Order No. 14154, 90 Fed. Reg. 8353 (Jan. 20, 2025).

292. Some courts have cast doubt on whether the GHG impacts of smaller regulatory decisions and agency actions are sufficiently large to establish standing for specific plaintiffs. See, e.g., *Wash. Env’t Council v. Bellon*, 732 F.3d 1131, 1135 (9th Cir. 2013) (emissions from Washington power plants amounting to 6% of state’s total GHG emissions not a “meaningful contribution” to climate change), *reh’g en banc denied*, 741 F.3d 1075 (9th Cir. 2014); *Amigos Bravos v. U.S. Bureau of Land Mgmt.*, 816 F. Supp. 2d 1118, 1136 (D.N.M. 2011) (254,730 metric tons of GHGs per year that might result from the approval of 92 oil and gas leases were not a “meaningful contribution” to global climate change).

293. For example, to support tort claims, plaintiffs may need to demonstrate that defendants made a “substantial contribution” to their injury (although the precise standard depends on the jurisdiction). Under the Clean Air Act, EPA has a legal obligation to regulate air pollution that will “cause or contribute” to the endangerment of public health and welfare. Under NEPA, the federal government has a legal obligation to conduct a comprehensive analysis of “significant” environmental impacts. These vague textual standards have been fleshed out, to some extent, through litigation—but as with standing, there are no firm quantitative thresholds for such standards, and these are ultimately subjective, case-specific determinations.

294. For example, many communities are already experiencing frequent floods and inundation as a result of sea-level rise. William Sweet et al., *Sea Level Rise and Nuisance Flood Frequency*

is potentially more difficult to downscale impact attribution claims to the level of individual injury, but this depends on the specific injury at issue.²⁹⁵

There are some cases where litigants do not need to demonstrate particularized harm as a result of climate change, but it is necessary for a plaintiff or defendant to establish a link between climate change and more generalized harm. This may be the case when a government agency is defending GHG regulations or other climate policies in court.²⁹⁶ For example, in *Coalition for Responsible Regulation v. EPA*, the D.C. Circuit Court of Appeals was tasked with evaluating the legality of a Clean Air Act endangerment finding in which EPA had determined that GHG emissions from motor vehicles “cause, or contribute to, air pollution which may reasonably be anticipated to endanger public health or welfare.”²⁹⁷ One critical question was whether EPA could rely on IPCC, USGCRP, and NRC reports to establish a causal link between motor vehicle GHG emissions and public endangerment. The court held that EPA’s reliance on these external studies was entirely proper and that the endangerment finding was supported by a “substantial” body of evidence amassed through decades of scientific research on climate change.²⁹⁸

The court specifically confronted arguments that there was “too much uncertainty” about the science underpinning climate change, and that EPA had improperly “delegated” its judgment to the IPCC, USGCRP, and NRC by relying on those assessments of climate science.²⁹⁹ The court disagreed on both

Changes Around the United States, NOAA Tech. Rep., NOS CO-OPS 073 (2014), <https://perma.cc/F4SM-BERS>; Kristina A. Dahl et al., *Effective Inundation of Continental United States Communities with 21st Century Sea Level Rise*, 5 *Elementa: Sci. of the Anthropocene* 1 (2017), <https://doi.org/10.1525/elementa.234>. Many communities have also been adversely affected by extreme events that were likely exacerbated by climate change, such as Hurricane Harvey. See Wehner & Samson, *supra* note 87.

295. The court in *Juliana v. United States*, 339 F. Supp. 3d 1062, 1096 (D. Or. 2018), *rev’d and remanded*, 947 F.3d 1159 (9th Cir. 2020), provides an example of how attribution research can factor into standing disputes as well as the merits of cases involving government obligations and individual rights. For a detailed discussion of that case, see Burger et al., *supra* note 77, at 164–66. See also Jessica Wentz, *Climate Science and Litigation*, in *Research Handbook on Climate Change Litigation* (Francesco Sindico et al. eds., 2024).

296. Another example would be NEPA lawsuits where plaintiffs are seeking to compel federal agencies to account for GHG impacts from federal actions. Although plaintiffs must establish standing in such cases, the causation requirements are relaxed for procedural claims under NEPA, and plaintiffs can also establish standing on the basis of injuries flowing from other deficiencies in the NEPA documentation. See, e.g., *WildEarth Guardians v. Jewell*, 738 F.3d 298 (D.C. Cir. 2013); *WildEarth Guardians v. Bureau of Land Mgmt.*, 8 F. Supp. 3d 17 (D.D.C. 2014). See also Jessica Wentz, *Environmental Impact Assessment*, in *Global Climate Change and U.S. Law* (Michael B. Gerrard et al. eds., 3d ed. 2023); Wentz, *supra* note 295; *Massachusetts v. EPA*, 549 U.S. 497, 498 (2007) (holding that standing requirements are relaxed where plaintiffs are seeking to enforce procedural rights created by Congress); *TransUnion LLC v. Ramirez*, 594 U.S. 413, 414 (2021) (holding that, even in the context of a statutory violation, the asserted injury must have a “close historical or common-law analogue”).

297. *Coal. for Responsible Regul. v. EPA*, 684 F.3d 102 (D.C. Cir. 2012); Clean Air Act § 202(a)(1), codified as amended at 42 U.S.C. 7521(a)(1).

298. *Coal. for Responsible Regul.*, 684 F.3d at 120.

299. *Id.* at 119–21.

counts. In upholding EPA’s reliance on the external studies, it noted that these “peer-reviewed assessments synthesized thousands of individual studies on various aspects of greenhouse gases and climate change” and that EPA was “not required to re-prove the existence of the atom every time it approaches a scientific question.”³⁰⁰ The court also rejected petitioners’ argument that EPA could not rely on IPCC assessments because some of the studies referenced therein were not peer reviewed—it noted that the IPCC assessment relied on approximately 18,000 studies that *were* peer reviewed, and petitioners had not “uncovered a ‘pattern’ of flawed science” in the assessments that undermined the endangerment finding.³⁰¹ On the more general issue of whether there was too much uncertainty about climate change impacts to justify the endangerment finding, the court found that EPA had supported its determination through three key lines of evidence: (1) “basic physical understanding” of the greenhouse effect, (2) observational evidence of past climate change, and (3) models predicting how the climate will respond to GHG concentrations in the future.³⁰²

Foreseeability

There are many different types of climate cases that implicate questions about the foreseeability of climate change impacts. Such questions tend to be prominent in cases involving adaptation and disclosure-related claims, where the core factual dispute may center on whether certain risks or impacts are “reasonably foreseeable” or “speculative,” as this is relevant to determining whether a defendant is required or authorized to take some action or make a disclosure in response to climate change. For example, disputes about the foreseeability of impacts are common in NEPA and ESA litigation. Questions about foreseeability may also arise in the context of causation inquiries—for example, where plaintiffs seek to establish standing based on an “imminent” injury, or when a government agency defends regulatory action based on the risk of future harm.

Both attribution research and predictive research can be used to evaluate the foreseeability of climate change impacts. Attribution research provides insights on the trends and impacts that are already occurring as a result of anthropogenic climate change, thus allowing a court to determine whether specific trends and impacts are hypothetical or speculative. Predictive research provides insights on the future trajectory of those trends and impacts under different emissions and warming scenarios. Courts can refer to scientific assessments of confidence and likelihood when evaluating whether specific trends and impacts are foreseeable as opposed to speculative or hypothetical.

300. *Id.* at 120.

301. *Id.* at 125.

302. *Id.* at 120–21.

Some of the most detailed scientific disputes involving the foreseeability of climate impacts have arisen in the context of ESA litigation.³⁰³ Many of the ESA cases address how federal agencies should deal with scientific uncertainty when evaluating the foreseeability and likelihood of specific trends and impacts that may pose a threat to a species (e.g., warmer ocean temperatures, the loss of Arctic sea ice, declines in food sources). Although courts are generally deferential to agency conclusions about climate change and scientific uncertainty,³⁰⁴ courts have also held that an agency cannot simply ignore climate-change-related threats on the basis of uncertainty or imprecision—the agency must formulate a rationale finding about the probability of threats, taking into account the “best scientific evidence,” including available climate research.³⁰⁵

There are also cases where parties may dispute the foreseeability of GHG emissions. Such disputes may arise, for example, in the context of NEPA litigation where plaintiffs are seeking disclosure of emissions from federal actions, and in the context of lawsuits involving corporate climate disclosures. In this context, courts can refer to source attribution data to determine whether emissions are reasonably foreseeable and should be addressed in NEPA documentation or other planning and disclosure documents.

Legal Obligations and Authorities

Climate science can also inform judicial determinations about the scope of a defendant’s legal obligations and/or authorities with regard to climate change mitigation, disclosure, and adaptation. Where climate science factors into such determinations, it is typically in relation to the analysis of causation and foreseeability, as discussed above.³⁰⁶ For example, whether a government agency is authorized or required to regulate GHG emissions from a particular source category may depend

303. See Daniel Kim et al., *Judicial Review of Scientific Uncertainty in Climate Change Lawsuits: Deferential and Nondeferential Evaluation of Agency Factual and Policy Determinations*, 46 *Harvard Env’t L. Rev.* 367 (2022), <https://perma.cc/YQ5E-BZY6>; Jessica Wentz, *Climate Change Attribution Science and the Endangered Species Act*, 39 *Yale J. Regul.* 1043 (2022), <https://perma.cc/YW2Y-QB4C>. Some significant scientific disputes have also arisen in the context of other wildlife protection statutes, such as the Marine Mammal Protection Act (MMPA). See, e.g., *Ctr. for Biological Diversity v. Kempthorne*, 588 F.3d 701 (9th Cir. 2009).

304. See, e.g., *Nat. Res. Def. Council, Inc. v. Coit*, 597 F. Supp. 3d 73 (D.D.C. 2022); *Ctr. for Biological Diversity v. Lubchenco*, 758 F. Supp. 2d 945 (N.D. Cal. 2010). See also *Kempthorne*, 588 F.3d 701 (upholding federal government’s conclusions about climate-change-related threats to species in the context of an MMPA claim).

305. See, e.g., *Ctr. for Biological Diversity v. Zinke*, 900 F.3d 1053 (9th Cir. 2018); *Greater Yellowstone Coal. Inc. v. Servheen*, 665 F.3d 1015 (9th Cir. 2011); *Def. of Wildlife v. Jewell*, 176 F. Supp. 3d 975 (D. Mont. 2016). See also Wentz, *supra* note 303; Kim, *supra* note 303.

306. This analysis would encompass questions related to whether emissions represent a material or substantial contribution to global climate change.

on whether there is sufficient record evidence that those emissions are contributing to public endangerment.³⁰⁷ Alternatively, whether a federal agency has an obligation to disclose GHG emissions and climate-change-related impacts under NEPA depends on whether those impacts are reasonably foreseeable, caused by the action under review, and sufficiently large such that they represent a material contribution to climate change.³⁰⁸ Federal agencies may also have NEPA obligations to disclose climate change impacts that are not caused by the action under review, specifically insofar as those impacts may affect baseline environmental conditions and influence the environmental outcomes of the project.³⁰⁹

The analysis of foreseeability also plays a prominent role in judicial assessments of legal obligations to assess, prepare for, and respond to the effects of climate change. For example, whether the federal government is required or authorized to issue protections for species under the ESA depends, in part, on whether there is a foreseeable risk of harm to those species. In ESA cases, courts have consistently held that climate science qualifies as the “best available science” that must be considered when agencies are making decisions about listing determinations, critical habitat designations, and other decisions involving the management and protection of endangered and threatened species.³¹⁰ Courts have also upheld the federal government’s use of climate science to support ESA protections for species that are at risk of extinction because of climate change, specifically rejecting arguments that projections of future climate change are “too speculative” to support the government’s assessment of risk.³¹¹

307. See, e.g., *Coal. for Responsible Regul. v. EPA*, 684 F.3d 102, 122 (D.C. Cir. 2012).

308. See, e.g., *Food & Water Watch v. Fed. Energy Regul. Comm’n*, 28 F.4th 277 (D.C. Cir. 2022); *Montana v. Haaland*, 50 F.4th 1254 (9th Cir. 2022); *Ctr. for Biological Diversity v. Bernhardt*, 982 F.3d 723 (9th Cir. 2020); *Sierra Club v. Fed. Energy Regul. Comm’n*, 867 F.3d 1357 (D.C. Cir. 2017); *Ctr. for Biological Diversity v. Nat’l Highway Traffic Safety Admin.*, 538 F.3d 1172 (9th Cir. 2008). For examples of cases where courts did not require disclosures for actions with smaller GHG impacts, see *Hapner v. Tidwell*, 621 F.3d 1239 (9th Cir. 2010); *Senville v. Peters*, 327 F. Supp. 2d 335 (D. Vt. 2004). See also Wentz (2023), *supra* note 296; Michael Burger & Jessica Wentz, *Evaluating the Effects of Fossil Fuel Supply Projects on Greenhouse Gas Emissions and Climate Change Under NEPA*, 44 Wm. & Mary Env’t L. & Pol’y Rev. 423 (2020), <https://perma.cc/P87F-RTYL>.

309. See, e.g., *AquAlliance v. U.S. Bureau of Reclamation*, 287 F. Supp. 3d 969, 1023–24 (E.D. Cal. 2018); *Conservation Cong. v. U.S. Forest Serv.*, No. 2:13-CV-01977-JAM-DB, 2018 WL 1142199 (E.D. Cal. Mar. 2, 2018).

310. See Alejandro E. Camacho, *Endangered Species Act*, in *Global Climate Change and U.S. Law* 237–54 (Michael B. Gerrard et al. eds., 2023); Wentz (2022), *supra* note 303. See also *Greater Yellowstone Coal., Inc. v. Servheen*, 665 F.3d 1015 (9th Cir. 2011); *Def. of Wildlife v. Jewell*, 176 F. Supp. 3d 975 (D. Mont. 2016); *Nat’l Res. Def. Council v. Kempthorne*, 506 F. Supp. 2d 322 (E.D. Cal. 2007); *Pac. Coast Fed’n Fishermen’s Ass’ns v. Gutierrez*, 606 F. Supp. 2d 1122, 1184 (E.D. Cal. 2008); *S. Yuba River Citizens League v. Nat’l Marine Fisheries Serv.*, 723 F. Supp. 2d 1247, 1274 (E.D. Cal. 2010); *Ctr. for Biological Diversity v. Salazar*, 804 F. Supp. 2d 987, 1008 (D. Ariz. 2011).

311. See, e.g., *Alaska Oil & Gas Ass’n*, 722 F. App’x 666 (9th Cir. 2018); *Alaska Oil & Gas Ass’n v. Pritzker*, 840 F.3d 671 (9th Cir. 2016); *In re Polar Bear Endangered Species Act Listing & 4(d) Rule Litig.*, 709 F.3d 1 (D.C. Cir. 2013).

Acknowledgments

The authors would like to acknowledge the insights and helpful feedback provided by Vivien Gornitz, Michael Gerrard, Michael Burger, and the committee responsible for development of the *Reference Manual on Scientific Evidence*.

Glossary of Terms

The following terms and definitions were adapted from various sources, including the Intergovernmental Panel on Climate Change (IPCC), the World Meteorological Organization (WMO), the U.S. Global Change Research Program (USGCRP), and the U.S. Environmental Protection Agency (EPA).

adaptation. The process of adjustment to actual or anticipated climate change and its effects.

anthropogenic climate change. A change in climate that can be attributed directly or indirectly to human activity.

anthropogenic emissions. Emissions of greenhouse gases (GHGs), precursors to GHGs, and aerosols caused by human activities. These activities include the burning of fossil fuels, deforestation, land use and land-use changes, livestock production, fertilization, waste management, and industrial processes.

atmosphere. The gaseous envelope surrounding the Earth that comprises five layers (troposphere, stratosphere, mesosphere, thermosphere, and exosphere). The atmosphere contains various gases, including water vapor and GHGs. The atmosphere is one component of the global climate system.

attribution. The process of evaluating the relative contributions of multiple causal factors to a change or event.

biogeochemical cycle. The biological, geological, and chemical cycling of elements such as carbon between different global systems (including living and nonliving systems). The carbon cycle is an example of a biogeochemical cycle.

biosphere. The part of the earth system comprising all ecosystems and living organisms, including within the atmosphere, on land (terrestrial biosphere), or in oceans (marine biosphere).

carbon budget. The amount of cumulative net global anthropogenic emissions that would result in limiting global warming to a given level within a given probability, taking into account the effect of other anthropogenic climate forcers. The carbon budget may be calculated in reference to carbon dioxide (CO₂) emissions exclusively, or in reference to combined emissions of CO₂, methane (CH₄), nitrous oxide (N₂O), and other gases.

carbon dioxide (CO₂). A heat-trapping gas. It occurs naturally but is also a by-product of burning fossil fuels and biomass, land-use change, and industrial processes. It is the principal anthropogenic GHG that affects the Earth's radiative balance. It is the reference gas against which the potency of other GHGs is measured and therefore has a global warming potential (GWP) of 1.

carbon sink. A type of carbon pool (natural or human, in soil, ocean, and plants) that stores or takes up more carbon than it releases into the atmosphere.

climate. The climate can be defined, alternatively, as the state of the climate system, or the average weather conditions over a period of time ranging from months to thousands or millions of years. The World Meteorological Organization (WMO) uses a standard period of thirty years to characterize climate conditions. See also climate system.

climate change. A change in the state of the climate (or global climate system) that can be identified by changes in the mean and/or variability of properties and that persists for an extended period, typically decades or longer. The climate system changes in time under the influence of its own internal dynamics and because of external forcings, such as volcanic eruptions, solar variations, orbital forcing, and anthropogenic forcings.

climate forcer. Any substance or process that affects the flow of energy coming into or out of the global climate system, thus affecting the amount of heat retained within the system. Anthropogenic climate forcers include GHGs, aerosols, and changes in land use that make land reflect more or less solar energy. There are also natural climate forcers, such as solar radiation and the particulate matter from volcanic eruptions. See also radiative forcing.

climate forcing. The net effect of various climate forcers on the flow of energy coming into or out of the global climate system.

climate model. A qualitative or quantitative representation of the climate system based on the physical, chemical, and biological properties of its components, as well as their interactions and feedback processes. Climate models are used as research tools to study and simulate the effect of anthropogenic forcings on the climate system.

climate projection. A simulated response of the climate system to a scenario of future emissions or concentrations of greenhouse gases, aerosols, and changes in land use, generally derived using climate models.

climate system. The global system that determines climate conditions, consisting of five major components (the atmosphere, the hydrosphere, the cryosphere, the lithosphere, and the biosphere) and the interactions among them.

confidence interval. A range of values calculated from the results of a study within which the true value is likely to fall. The range of this interval depends on the confidence level—e.g., if the confidence level is 95%, then there is a 95% probability that the true value of a parameter will fall within the range of values provided in the confidence interval. The size of a confidence interval provides insight on the level of precision and range of uncertainty in study results.

confidence level. The probability that a particular value falls within a specified confidence interval.

cryosphere. The frozen components of the earth system, including sea ice, lake ice, river ice, snow cover, glaciers, ice sheets, ice caps, ice shelves, and frozen ground (permafrost). The cryosphere is part of the global climate system.

detection. The process of demonstrating that the climate system (or a system affected by the climate system) has changed in some defined statistical sense. As per IPCC definitions, an identified change is detected in observations if its likelihood of occurrence because of internal variability alone is determined to be small, for example $<10\%$.

fraction of attributable risk. In climate change research, the fraction of attributable risk (FAR) is the ratio of a particular outcome occurring with and without human influence on climate. Mathematically speaking, $FAR = 1 - P0/P1$ (where $P1$ equals the probability of the outcome occurring in the presence of anthropogenic forcing of the climate system, and $P0$ equals the probability of the event occurring if the anthropogenic forcing were not present).

global warming potential (GWP). A measurement of the radiative forcing from a unit of a given substance, accumulated over a chosen time horizon, relative to that of CO_2 .

greenhouse gas (GHG). A gas that absorbs incoming solar radiation (i.e., heat) in the atmosphere, preventing it from escaping into space. GHGs in the Earth's atmosphere include water vapor, carbon dioxide (CO_2), methane (CH_4), nitrous oxide (N_2O), ozone, chlorofluorocarbons (CFCs), hydrochlorofluorocarbons (HCFCs), hydrofluorocarbons (HFCs), perfluorocarbons (PFCs), sulfur hexafluoride (SF_6), and nitrogen trifluoride (NF_3).

hydrosphere. The components of the earth system including liquid surface and subterranean water, such as oceans, seas, rivers, freshwater lakes, underground water, and wetlands. The hydrosphere is part of the global climate system.

lithosphere. The rocky outer portion of the Earth, composed of the crust and upper part of the mantle. The lithosphere is part of the climate system, as it absorbs and reflects solar energy, radiates heat, stores carbon, and interacts with regional climate conditions.

methane (CH_4). A greenhouse gas; methane is the main constituent of natural gas, and key anthropogenic sources of methane include oil and gas systems, landfills, agricultural and livestock activities, coal mining, fossil fuel combustion, wastewater treatment, industrial processes, and land use changes. Methane is more than twenty-five times as potent as CO_2 , but it has a shorter atmospheric lifetime and is eventually converted into CO_2 in the atmosphere. Methane is the second largest contributor to anthropogenic climate forcing, after CO_2 .

model ensembles. Can refer to both (1) the full set of models, for example climate models developed by different modeling centers, used by the IPCC and other bodies for risk assessment, and (2) the averaging of results across multiple such models. The averaging of multiple models is generally considered more reliable than the use of a single model, although presenting the results separately from all the models is often considered superior to showing only the average across models, since the former gives a sense of the range of possible outcomes and important uncertainties.

natural variability. Naturally occurring variations in the mean state and other statistical measures of the climate.

nitrous oxide (N₂O). A greenhouse gas produced by agriculture, fuel combustion, industrial processes, and waste management systems, as well as natural systems. Nitrous oxide is approximately 300 times more potent than CO₂ but it is shorter lived in the atmosphere. Nitrous oxide also contributes to ozone depletion. It is the third largest contributor to anthropogenic climate forcing, after CO₂ and methane.

radiative forcing. The change in energy flux within the Earth's atmosphere caused by natural or anthropogenic climate forcers (e.g., increasing atmospheric concentrations of CO₂). Positive radiative forcing occurs when the Earth receives more incoming energy from sunlight than it radiates into space, and this net gain of energy causes warming.

relative risk. In climate studies, the relative risk is the ratio of probabilities for an event or impact occurring with and without anthropogenic influence on the climate system (or at varying levels of anthropogenic influence on the climate system).

tipping point. A critical threshold beyond which a system reorganizes, often abruptly and/or irreversibly.

References

Select References on Climate Science

- American Meteorological Society, Explaining Extreme Events from a Climate Perspective, <https://www.ametsoc.org/ams/index.cfm/publications/bulletin-of-the-american-meteorological-society-bams/explaining-extreme-events-from-a-climate-perspective>.
- Climate Science*, in Stanford Encyclopedia of Philosophy (first published May 11, 2018), available at <https://plato.stanford.edu/entries/climate-science>.
- IPCC, Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the IPCC (2021).
- IPCC, Climate Change 2022: Impacts, Adaptation, and Vulnerability, Working Group II Contribution to the Sixth Assessment Report of the IPCC (2022).
- IPCC, Climate Change 2022: Mitigation of Climate Change, Working Group III contribution to the Sixth Assessment Report of the IPCC (2022).
- IPCC, Global Warming of 1.5°C. An IPCC Special Report on the Impacts of Global Warming of 1.5°C Above Pre-Industrial Levels and Related Global Greenhouse Gas Emission Pathways (2018).
- National Academies of Science, Engineering, and Medicine, Attribution of Extreme Weather Events in the Context of Climate Change (2016).
- National Research Council, Climate Change: Evidence and Causes: Update 2020 (2020).
- U.S. Global Change Research Program, Fourth National Climate Assessment, Vol 1: Climate Science Special Report (2017).
- U.S. Global Change Research Program, Fifth National Climate Assessment (2023).
- U.S. Global Change Research Program, Second State of the Carbon Cycle Report (2018).
- WMO, Provisional State of the Global Climate 2022 (Nov. 2022).

Select References on Climate Science and the Law

- Myles Allen, *Liability for Climate Change*, 421 Nature 891 (2003).
- Maria L. Banda, *Climate Science and the Courts: A Review of U.S. and International Judicial Pronouncements*, Environmental Law Institute (2020).
- Michael Burger et al., *The Law and Science of Climate Change Attribution*, 5 Columbia Journal of Environmental Law 57 (2020).
- Daniel Kim et al., *Judicial Review of Scientific Uncertainty in Climate Change Lawsuits: Deferential and Nondeferential Evaluation of Agency Factual and Policy Determinations*, 46 Harvard Environmental Law Review 367 (2022).
- E.A. Lloyd & T.G. Shepherd, *Climate Change Attribution and Legal Contexts: Evidence and the Role of Storylines*, 167(3–4) Climatic Change 28 (2021).

Appendix: Overview of Scientific Organizations and Government Agencies Involved in the Production, Synthesis, and Dissemination of Climate Science

World Meteorological Organization (WMO)

The WMO is a United Nations (UN) agency responsible for coordinating and advancing climatological and meteorological research around the world.³¹² One key goal of the WMO is to promote the “free and unrestricted” exchange of data, information, and research between the 193 countries and territories that are signatories to the WMO convention.³¹³ The WMO is governed by the World Meteorological Congress, which is composed of member states.

Although the WMO is a political organization, it is recognized as an authoritative source of climatological and meteorological data. It provides standardized data products for climate research, most notably the “Climatological Standard Normals,” which provide information on typical climate conditions averaged over thirty-year periods for thousands of locations across the world.³¹⁴ The WMO also publishes an annual *State of the Global Climate Report* that summarizes the latest observations and findings for various “global climate indicators” including greenhouse gases, global temperature, ocean heat content, sea level, marine heat waves, the cryosphere, and precipitation.³¹⁵ The report also discusses natural phenomena that are important drivers of year-to-year variability in weather patterns, such as the El Niño Southern Oscillation, and explains how these have influenced recent weather observations.

Intergovernmental Panel on Climate Change (IPCC)

The IPCC is widely considered to be the leading scientific body for the assessment and synthesis of research on climate change. It was established in 1988 by the United Nations Environment Programme (UNEP) and the WMO as an

312. The WMO originated from the International Meteorological Organization (IMO), which was a nongovernmental organization founded in the late 1800s in order to promote the exchange of meteorological research. The WMO was subsequently established as a direct successor of the IMO through an agreement of UN member states in 1950.

313. WMO, *Overview*, <https://perma.cc/SC9L-CTSL>.

314. See WMO, *WMO Climatological Normals*, <https://perma.cc/65TT-69XV>.

315. See, e.g., WMO, *Provisional State of the Global Climate*, <https://perma.cc/X8J7-KVX7>.

intergovernmental body of the UN. Its primary mandate is to provide governmental bodies and the public with regular assessments of the scientific basis for climate change, its impacts and future risks, and options for mitigation and adaptation.³¹⁶ It is governed by 195 member states.³¹⁷ There are also nearly 200 observer organizations that participate in IPCC activities—for example, by nominating scientists to serve as report authors, and by reviewing draft reports.³¹⁸

The IPCC publishes periodic assessments of the state of the scientific knowledge on anthropogenic climate change. The IPCC does not conduct its own research; rather, it conducts systematic reviews of relevant scientific publications and then synthesizes key findings from the literature in its assessments. It does so with the assistance of thousands of scientists with diverse expertise across the field of climate research.³¹⁹ There are three IPCC working groups involved in the assessment process: Working Group I synthesizes research on the physical-science basis for anthropogenic climate change, including research on the changes that are already underway as well as projections of future climate change; Working Group II synthesizes research on the physical, social, and economic impacts of climate change; and Working Group III synthesizes research on potential pathways for mitigating and adapting to climate change. The experts involved in these working groups are selected by IPCC member governments, observer organizations, and the IPCC Bureau.

The IPCC conducts a comprehensive assessment of climate science every six to seven years. During this process, each working group prepares a full “assessment report” that contains a detailed technical discussion of its respective topic.³²⁰ Each draft report undergoes an initial review by a team of scientific and technical experts and then a second review by governments and experts. A

316. See IPCC, History of the IPCC, <https://perma.cc/W5K7-PT4B>.

317. Representatives of IPCC member governments meet one or more times a year in plenary sessions. The member states elect a Bureau of Scientists to guide technical activities for each assessment cycle. They also review and approve draft assessments. IPCC, Structure of the IPCC, <https://perma.cc/7LZ2-F9CP>.

318. Any nonprofit body or agency with relevant expertise may apply to become an IPCC observer organization. The current list of observer organizations includes UN bodies and organizations, intergovernmental organizations, government agencies, universities, scientific research organizations, environmental NGOs, and industry NGOs. Representatives of observer organizations may attend sessions of the IPCC and the plenary sessions of the IPCC working groups. They are also invited to encourage experts to review draft IPCC reports. See IPCC, *Structure*, *supra* note 317; IPCC, Observer Organizations, <https://perma.cc/697T-GKC7>.

319. The scientists involved in the assessments are selected through a nomination process. Prior to initiating an assessment, the IPCC issues a call to governments and IPCC observer organizations for nominations; the authors are then selected by the Bureau of Scientists on the basis of their expertise. The IPCC seeks to build author teams that reflect a range of scientific, technical, and socioeconomic expertise. See IPCC, Factsheet: How Does the IPCC Select Its Authors?, <https://perma.cc/45V6-W8TB>.

320. For a complete explanation of the process governing IPCC assessments, see IPCC, Procedures, <https://perma.cc/C39M-TW6H>.

selection of scientists within each working group then produces a Summary for Policymakers (SPM), which is intended to help policy makers and nonscientists understand key findings from each report. The IPCC member states conduct a detailed review of each SPM, in consultation with scientists, and they must approve the SPM line-by-line. The IPCC member states also approve the full reports, but that approval process is less detailed and government inputs are more limited (as compared with the SPMs). At the end of the assessment process, the IPCC also prepares a Synthesis Report that combines findings from all three working groups. The Synthesis Report goes through the same type of review process as other reports, with an SPM subject to detailed review by member states.

The IPCC also publishes methodology reports, technical papers, and other special reports in response to requests for scientific and technical information from the United Nations Framework Convention on Climate Change (UNFCCC), governments, and international organizations. For example, in 2018, the IPCC issued a special report evaluating the impacts of climate change if anthropogenic emissions caused global warming at or in excess of 1.5°C.³²¹

Although the IPCC is generally considered to be a global authority on climate science, it has been critiqued on the grounds that it is a political institution subject to the governance of member states. Concerns have also been raised about the inclusion of experts with potential bias in the assessment process (e.g., representatives from renewable energy industry associations or, conversely, from fossil fuel industry associations).³²² While no organization is perfect, an argument can be made that the structure of the IPCC actually contributes to institutional legitimacy and helps mitigate individual bias on the part of report authors and government representatives.³²³ The assessments are prepared with the help of a large number of contributors, including the report authors (which include physical scientists as well as researchers with other expertise), governmental representatives, and observer organization representatives, and they reflect the consensus on the part of both scientific experts and governments. The assessments have also received criticism from both sides of the climate change debate (some arguing that they overestimate the severity of anthropogenic climate change, others arguing that they underestimate the severity).³²⁴ This suggests that there is not a pervasive industry or environmental bias within the IPCC.

321. IPCC 1.5°C Report, *supra* note 253, at 5.

322. See, e.g., Editorial: *Evolving the IPCC*, 1 *Nature Climate Change* 227 (2011), <https://doi.org/10.1038/nclimate1189>.

323. See, e.g., Eric Paglia & Charles Parker, *The Intergovernmental Panel on Climate Change: Guardian of Climate Science*, in *Guardians of Public Value* (A. Boin et al. eds., 2021).

324. See, e.g., Naomi Oreskes, *Climate Change Will Cost Us Even More Than We Think*, N.Y. Times, Oct. 23, 2019.

Other Major Scientific Organizations

There are several U.S.-based organizations that play a key role in the dissemination of climate research. These include the American Meteorological Society (AMS), the American Geophysical Union (AGU), and the National Academies of Sciences, Engineering, and Medicine (National Academies or NASEM). The scientific community views all three as credible sources of climate science.

The National Academies were established by the federal government in 1863, with a specific mandate of providing scientific assessments to help support government decision-making. Although the National Academies' mandate is to assist the federal government, they are fully independent from the government.³²⁵ Like the IPCC, the National Academies do not conduct independent research; rather they assess and synthesize existing research in order to provide objective scientific advice. They periodically publish reports on scientific matters of national significance. They have published numerous reports on the topic of climate change, including *Climate Change: Evidence and Causes*³²⁶ and *Attribution of Extreme Weather Events in the Context of Climate Change*.³²⁷ The National Academies also maintain peer-reviewed scientific journals, most notably the *Proceedings of the National Academies of Sciences* (PNAS), and they are involved in reviewing the *National Climate Assessments* prepared by the federal government (discussed below). For many years, the National Research Council (NRC) was the operating arm of the National Academies (e.g., the NRC published many scientific reports); but the NRC has since been reorganized into seven program units with different thematic focuses.³²⁸

The AMS is a professional membership organization involved in the dissemination of atmospheric, oceanic, and hydrologic sciences. It was founded in 1919 with the mission of advancing scientific research in these fields, and since then it has grown to a membership of nearly 12,000 scientific professionals.³²⁹ The AMS maintains a number of publications that disseminate new research in climate science and related fields. These include the *Bulletin of the American Meteorological Society* (BAMS), the *Journal of the Atmospheric Sciences*, the *Journal of Applied Meteorology and Climatology*, *Earth Interactions*, and the *Journal of Climate*, among others. These are all peer-reviewed publications with excellent standing in the scientific community. Each year, AMS also publishes a *State of the Climate* report as a supplement to BAMS as well as a special edition of BAMS on *Explaining Extreme Events from a Climate Perspective*.³³⁰ This is an annual review of

325. NASEM, About Us, <https://perma.cc/972L-WJPK>.

326. NAS Update, *supra* note 6.

327. NASEM, *supra* note 91, at 51.

328. NASEM, Our Program Divisions and Units, <https://perma.cc/ZXZ4-MXGG>.

329. AMS, About the American Meteorological Society, <https://perma.cc/Y4MA-CXCW>.

330. BAMS, Explaining Extreme Events from a Climate Perspective, <https://perma.cc/ZG2C-LRZS>.

peer-reviewed research assessing how anthropogenic climate change has affected the magnitude, likelihood, and other characteristics of extreme events such as floods and heat waves.

The AGU is a professional membership organization involved in the dissemination of atmospheric, ocean, hydrologic, space, and planetary research. It was established in 1919 by the NRC and it operated as an affiliate of the National Academies until it was independently incorporated in 1972.³³¹ It now has more than half a million members from around the world. Like AMS, AGU maintains various peer-reviewed publications that disseminate new research in climate science and related fields. These include the *Journals of Geophysical Research*, the *Journal of Advances in Modeling Earth Systems*, *Earth and Space Science*, and *Geophysical Research Letters*. As with AMS publications, the AGU journals have excellent standing in the scientific community.

Government Agencies

The U.S. federal government also plays a major role in both the production and dissemination of climate research.³³² The federal programs and agencies that have been most heavily involved in the production of climate science outputs include the U.S. Global Change Research Program (USGCRP), the National Oceanic and Atmospheric Administration (NOAA), the National Aeronautics and Space Administration (NASA), the U.S. Geological Survey (USGS), and the Environmental Protection Agency (EPA).

The USGCRP has been the leading federal entity involved in the coordination and dissemination of research on climate change. The program began as a presidential initiative in 1989 and was subsequently codified by Congress in 1990 as “a comprehensive and integrated United States research program which will assist the Nation and the world to understand, assess, predict, and respond to human-induced and natural processes of global change.”³³³ Per this Congressional directive, the USGCRP has coordinated federal research on climate change and has periodically published *National Climate Assessments* (NCAs), which

331. AGU, Learn About AGU, <https://perma.cc/RFY9-6WZ3>.

332. Many states also have government agencies involved in the production and dissemination of climate research. But for the purposes of this reference guide, we focus on scientific outputs from federal agencies, as federal judges are more likely to engage with these outputs.

333. Global Change Research Act of 1990, Pub. L. 101–606, title I, § 101, 104 Stat. 3096 (codified at 15 U.S.C. § 2931). Presidential Administrations and Congress have taken various positions on the program. Kathryn G. Kynett, Cong. Rsch. Serv., R48478, U.S. Global Change Research Program (USGCRP): Overview and Considerations for Congress (2025). In April 2025, several federal contracts supporting USGCRP were reportedly canceled and federal staff previously involved in the program were reportedly dismissed. At the time of writing, the USGCRP website appears to no longer be available. *See id.*

summarize the latest science on anthropogenic climate change, including research on how climate change is affecting and may affect the United States now and in the future.³³⁴ These reports have been a useful supplement to IPCC assessments insofar as they focus on the U.S. context. The *Fifth National Climate Assessment* (NCA5) was published in 2023.

A number of other federal agencies, including NOAA, NASA, USGS, and EPA, are directly involved in climate research and the production of climate data products.³³⁵ NOAA plays a particularly important role in federal climate research and dissemination—it maintains records of climate indicators (e.g., national precipitation and temperature records) and publishes monthly reports summarizing the latest climate data for the United States and the globe.³³⁶ NASA also produces and maintains climate data products—these include satellite data as well as outputs from NASA weather and climate models.³³⁷ The USGS carries out geophysical research on climate change and its impacts as part of its Ecosystems Land Change Science Program,³³⁸ and it also maintains national and regional Climate Adaptation Science Centers, which are collaborative scientific partnerships aimed at building knowledge to help resource managers and communities prepare for and respond to the effects of climate change.³³⁹ EPA primarily conducts research on the environmental and public health effects of climate

334. Among other responsibilities, the USGCRP is required by Congress to develop an assessment of global change at least every four years, which means that the deadline to submit the Sixth National Climate Assessment to Congress is 2027. See 15 U.S.C. § 2931. A court enforced this deadline in the past, in *Center for Biological Diversity v. Brennan*, 571 F. Supp. 2d 1105 (N.D. Cal. 2007).

335. This is not an exhaustive list of all federal entities involved in climate research. So many agencies are involved in climate research because climate change research involves many areas of science and has such far-reaching implications for the U.S. government, citizens, and natural resources. Other departments and agencies, including but not limited to the Department of Energy (DOE), the Department of Interior (DOI), the Department of Defense (DOD), the Department of Agriculture (USDA), and agencies and services within these departments, also conduct climate-related research relevant to their respective mandates. Recently, the DOE established a Climate Working Group, consisting of five scientists assembled by the DOE Secretary. The DOE released a report written by this group, which was still in draft form at the time of this writing, in tandem with EPA's proposed rule repealing the 2009 endangerment finding. See U.S. Dep't of Energy, Climate Working Group, A Critical Review of Impacts of Greenhouse Gas Emissions on the U.S. Climate (July 23, 2025), <https://perma.cc/KHF8-TTQR>; Dep't of Energy, Climate, <https://perma.cc/647D-7224> (last visited Oct. 8, 2025) (providing a report evaluating the impact of greenhouse gasses on U.S. climate and inviting the public to comment). See also *supra* note 290 and accompanying text for a related discussion of the endangerment finding.

336. NOAA, National Centers for Environmental Information: Products by Category, <https://perma.cc/S9U9-V7W6>.

337. NASA, Data Resources, <https://perma.cc/TB99-BJA7>; NASA Center for Climate Simulation, Climate Data Services, <https://perma.cc/72FX-U5EZ>.

338. USGS, Ecosystems Land Change Science Program, <https://perma.cc/N67D-3TQ3>.

339. USGS, Climate Adaptation Science Centers, <https://perma.cc/6W9Y-736D>.

change—for example, examining how climate change will affect air quality and ecosystems.³⁴⁰

Nongovernmental Organizations

In addition to the international and U.S.-based organizations described above, there are many nongovernmental and academic organizations involved in the production and dissemination of climate research and data products. Some examples include the World Weather Attribution project, a collaboration of climate scientists that provides rapid assessments on the role of anthropogenic climate change in extreme weather events;³⁴¹ the Climate Impact Lab, a coalition of scientists, economists, and other analysts that conduct research on the impacts and costs of climate change;³⁴² and Climate Signals, an initiative that provides written synthesis of and data visualization products for climate attribution research.³⁴³

Many of these entities provide high-quality research products. However, these organizations do not have the same scientific standing as the major scientific organizations discussed above, and their research products do not always undergo formal peer review. There are various ways in which to evaluate the credibility of these research products. The credentials of the individuals producing the products are very important—the World Weather Attribution project, for example, consists of scientific researchers at top universities who have extensive experience in the climate and atmospheric sciences. Funding sources may also provide insights on credibility—for example, if a group is funded by fossil fuel interests or other industry actors that have a financial stake in preventing the regulation of fossil fuels, this could raise red flags about the credibility of any research findings. Finally, research products can also be evaluated for consistency with climate assessments from authoritative institutions like the IPCC.

340. EPA, Climate Change Research, <https://perma.cc/HQH9-XZWH>.

341. World Weather Attribution, <https://perma.cc/6W6F-VRVG>.

342. Climate Impact Lab, <https://perma.cc/FK4H-3Z7Y>.

343. Climate Signals, <https://perma.cc/U8MG-927K>.

Appendix

Biographical Information of Committee and Staff

Co-Chairs

Nancy D. Freudenthal is a U.S. District Judge for the U.S. District Court for the District of Wyoming. Freudenthal was born and raised in Cody, WY, and earned both a bachelor's degree and a juris doctorate from the University of Wyoming. She worked for former Wyoming governors Ed Herschler and Mike Sullivan, and then was appointed to the State Board of Equalization. She entered private practice in 1995 with the law firm of Davis & Cannon. In 2010, following a presidential appointment and Senate confirmation, Judge Freudenthal became the first woman appointed to the federal court in Wyoming and the seventh federal district judge in the state's history.

Fred H. Gage, Ph.D., is the Adler Professor in the Laboratory of Genetics at the Salk Institute for Biological Studies and an adjunct professor at the University of California, San Diego. He served as president of the Salk Institute from 2018 to 2023. He received his Ph.D. from Johns Hopkins University. Gage's work concentrates on the adult central nervous system and unexpected plasticity and adaptability to environmental stimulation that remains throughout the life of all mammals. In addition, he studies human neurological and psychiatric disease in vitro using human stem cells. His lab also studies the genomic mosaicism that exists in the brain as a result of mobile elements that are active during neurogenesis. Gage is a fellow of the American Association for the Advancement of Science, a member of the National Academy of Sciences, the National Academy of Medicine, and the American Philosophical Society, a foreign member of the European Molecular Biology Organization, and a member of the American Academy of Arts and Sciences. He served as president of the Society for Neuroscience in 2002, and of the International Society for Stem Cell Research in 2012.

Members

Russ Biagio Altman is the Kenneth Fong Professor of Bioengineering, Genetics, Medicine, Biomedical Data Science, and (by courtesy) Computer Science at Stanford University, senior fellow at the Stanford Institute for HAI (Human-Centered

Artificial Intelligence), and past chairman of the Bioengineering Department at Stanford University. Altman's research applies computing (informatics, data science, and artificial intelligence) to problems relevant to medicine. He is interested in methods for understanding drug action, efficacy, and adverse effects while integrating molecular, cellular, organism, and population data. His lab has a special focus on how human genetic variation impacts drug response, or pharmacogenomics. He is the founding principal investigator of the Pharmacogenomics Knowledgebase (PharmGKB.org). Altman helps lead an FDA-supported Center of Excellence in Regulatory Science & Innovation and is an associate director of the Stanford Institute for HAI. Altman holds an A.B. from Harvard College, an M.D. from Stanford Medical School, and a Ph.D. in Medical Information Sciences from Stanford. He received the U.S. Presidential Early Career Award for Scientists and Engineers and a National Science Foundation CAREER Award. He is a fellow of the American College of Physicians (ACP), the American College of Medical Informatics (ACMI), the American Institute of Medical and Biological Engineering (AIMBE), and the American Association for the Advancement of Science (AAAS). He is a member of the National Academy of Medicine. He is a past-president, founding board member, and a fellow of the International Society for Computational Biology (ISCB), and a past-president of the American Society for Clinical Pharmacology & Therapeutics (ASCPT). He has chaired the Science Board advising the FDA commissioner, served on the NIH Director's Advisory Committee, and co-chaired the Institute of Medicine's Drug Forum. He is an organizer of the annual Pacific Symposium on Biocomputing, and a founder of Personalis. Altman is board certified in Internal Medicine and in Clinical Informatics. He directed the Stanford Biomedical Informatics training program from 2000–2018 and currently directs the undergraduate major in Biomedical Computation. Altman received the Stanford Medical School graduate teaching awards in 2000 and 2020 and its mentorship award in 2014. He received the Stanford Alumni Association Kornberg–Berg Lifetime Achievement Award in 2023. He is the founding editor of the *Annual Reviews of Biomedical Data Science*, and hosts a podcast titled *The Future of Everything*.

David G. Campbell is a Senior U.S. District Judge for the District of Arizona, having served on the court since 2003. Campbell currently chairs the U.S. Courts' Committee on International Judicial Relations. He previously chaired the U.S. Committee on Rules of Practice and Procedure, which oversees the work of the five Advisory Committees on the Federal Rules of Civil, Criminal, Bankruptcy, and Appellate Procedure and the Federal Rules of Evidence. He also chaired the Advisory Committee on the Federal Rules of Civil Procedure from 2011 to 2015 and served as a member of the committee from 2005 to 2011. Campbell is a member of the American Law Institute and a fellow of the American Bar Foundation. He has travelled to Botswana, Namibia, South Africa, Turkey, Gambia, Slovakia, South Korea, Bulgaria, the Philippines, and other countries on judicial case

Appendix: Biographical Information of Committee and Staff

management and rule-of-law projects and has taught civil procedure and constitutional law at the Arizona State and Brigham Young University Law Schools. Following his graduation from the University of Utah Law School, he served as a law clerk for Justice William H. Rehnquist of the U.S. Supreme Court and Judge J. Clifford Wallace of the U.S. Court of Appeals for the Ninth Circuit. Campbell worked as a commercial litigator with the Phoenix, Arizona, law firm of Osborn Maledon before becoming a judge.

Alicia L. Carriquiry is a Distinguished Professor of Liberal Arts and Sciences and the President's Professor of Statistics at Iowa State University. She researches applications of statistics in human nutrition, bioinformatics, forensic sciences, and traffic safety. Carriquiry has published more than 160 peer-reviewed articles in corresponding academic journals.

Carriquiry is the lead investigator for the Center for Statistics and Applications in Forensic Evidence (CSAFE) program, providing scientific oversight and research expertise. Along with Drs. Hal S. Stern and Michael Daniels, she was among the first to question the probative value of bullet lead analysis. A 2000 report to the Federal Bureau of Investigation that suggested the probability of a coincidental match might not be negligible led to the establishment of a National Academies of Sciences, Engineering, and Medicine (National Academies) committee to explore the issue. Carriquiry subsequently served as a member of the National Academies' Committee on Assessing the Feasibility, Accuracy, and Technical Capability of a National Ballistics Database. Overall, Carriquiry has chaired or participated in over 15 National Academies boards, panels and committees and is currently a statistics editor for the *Proceedings of the National Academy of Sciences*. Carriquiry has spoken at many national and international conferences in statistics, forensic science, and nutrition and has been invited to deliver multiple named lectures and seminars. She has served on several subcommittees in the National Institute for Standards and Technology (NIST) Organization of Scientific Area Committees for Forensic Science (OSAC) and is a member of the U.S. Technical Advisory Group to the International Organization for Standardization (ISO) Committee on Forensic Science. Carriquiry was elected to the National Academy of Medicine in 2016 and to the American Academy of Forensic Sciences in 2023. She is a fellow of the American Association for the Advancement of Science, the American Statistical Association, the Institute of Mathematical Statistics, the International Society for Bayesian Analysis, and the International Statistical Institute. She was recently elected fellow of the Hagler Institute for Advanced Studies at Texas A&M University. Carriquiry received an M.S. in animal science from the University of Illinois, and an M.S. and a Ph.D. in statistics and animal genetics from Iowa State University.

Lynn R. Goldman, a pediatrician and an epidemiologist, is the Michael and Lori Milken Dean and Professor of Environmental and Occupational Health at the

Milken Institute School of Public Health at the George Washington University. She was previously a professor of environmental health sciences at the Johns Hopkins Bloomberg School of Public Health; an assistant administrator for toxic substances at the U.S. Environmental Protection Agency, where she directed the Office of Chemical Safety and Prevention; and the chief of the Division of Environmental and Occupational Disease Control at the California Department of Public Health. Goldman completed a B.S. and M.S. from the University of California, Berkeley, an M.D. from the University of California, San Francisco (UCSF); an M.P.H. from the Johns Hopkins Bloomberg School of Public Health; and pediatric residency at the UCSF Benioff Children's Hospital in Oakland.

Goldman is a member of the National Academy of Medicine (NAM). She serves as a trustee of the Environmental Defense Fund; a member of the National Academy of Sciences, Engineering and Medicine's Environmental Health Matters Initiative; a member of the American Academy of Pediatrics Council on Environmental Health Executive Committee; cochair of the Association of Schools and Programs of Public Health Climate Change and Health Task Force; and a member of the National Academies of Sciences, Engineering, and Medicine's Committee on Managed Retreat in the U.S. Gulf Coast Region.

Brian W. Kernighan is the William O. Baker '39 Professor in Computer Science at Princeton University. He received his B.A.Sc. in engineering physics at the University of Toronto in 1964 and his Ph.D. in electrical engineering from Princeton in 1969. Kernighan was a member of technical staff and department head in the Computing Science Research Center at Bell Laboratories from 1969 until 2000 and was named a Bell Labs fellow in 1990. He joined the Princeton Department of Computer Science in 2000. Kernighan is a co-creator of several programming languages, including AWK and AMPL, and of a number of tools for document preparation. He holds five patents, and is a co-founder of AMPL Optimization Inc. Kernighan's research areas include programming languages, tools, and interfaces that make computers easier to use, often for non-specialist users. He is the co-author of a number of widely used computer science books, including *The C Programming Language* (with Dennis Ritchie) and *The Unix Programming Environment* (with Rob Pike). He has also written books on technology for non-technical audiences: *Understanding the Digital World*; *Millions, Billions, Zillions: Defending Yourself in a World of Too Many Numbers*; and *Unix: A History and a Memoir*. Kernighan is a member of the National Academy of Engineering and the American Association of Arts and Sciences.

Pramod P. Khargonekar, Ph.D., is Vice Chancellor for Research and Distinguished Professor of Electrical Engineering and Computer Science at the University of California, Irvine. His research and teaching interests are centered on confluence of machine learning and control; advanced manufacturing; energy

Appendix: Biographical Information of Committee and Staff

system decarbonization; climate change mitigation, adaptation, and resilience; and societal impacts of technology. Previously, after serving briefly as Deputy Director of Technology at the Advanced Research Projects Agency – Energy in 2012–13, Khargonekar served, until June 2016, as the head of the Directorate of Engineering at the National Science Foundation. From 2001 to 2009, he was Dean of the College of Engineering at the University of Florida and was the Eckis Professor of Electrical and Computer Engineering at the university until 2016. Khargonekar was chair of the Department of Electrical Engineering and Computer Science at the University of Michigan from 1997 to 2001, where he also held the position of Claude E. Shannon Professor of Engineering Science. At the University of Michigan, he received the Claude Shannon Chair and Arthur F. Thurnau Professorship. Khargonekar has been recognized as a Web of Science Highly Cited Researcher. He is a recipient of the IEEE (Institute of Electrical and Electronics Engineers) Control Systems Award, IEEE Control Systems Society Bode Lecture Prize, National Science Foundation Presidential Young Investigator Award, the American Automatic Control Council’s Donald Eckman Award, the Japan Society for Promotion of Science fellowship, World Automation Congress Honor, the IEEE W. R. G. Baker Prize Award, the IEEE CSS George Axelby Best Paper Award, the Hugo Schuck American Control Conference (ACC) Best Paper Award, and distinguished alumnus and distinguished service awards from the Indian Institute of Technology, Bombay. Khargonekar is an inaugural member of the Hall of Fame, Department of Electrical and Computer Engineering, University of Florida, and a fellow of IEEE, the International Federation of Automatic Control (IFAC), and the American Association for the Advancement of Science.

Goodwin Liu is an Associate Justice of the California Supreme Court. He was nominated by Governor Jerry Brown and sworn into office in 2011 and retained by the electorate in 2014 and 2022. Before joining the state’s highest court, Liu was Professor of Law and Associate Dean at the University of California, Berkeley School of Law. His areas of expertise span constitutional law, education law and policy, and diversity in the legal profession. Liu is an elected member and chair of the board of the American Academy of Arts & Sciences. He is also an elected member of the American Law Institute and American Philosophical Society, in addition to having served of the board of numerous service organizations and academic institutions. The son of Taiwanese immigrants, Liu grew up in Sacramento and earned a bachelor’s degree in biology from Stanford University, master’s degrees in philosophy and physiology from Oxford University on a Rhodes Scholarship, and J.D. from Yale Law School. He clerked for Judge David S. Tatel on the U.S. Court of Appeals for the D.C. Circuit and for Justice Ruth Bader Ginsburg on the U.S. Supreme Court. His accomplished career also includes practicing at O’Melveny & Myers, serving in the U.S. Department of Education, and helping launch the AmeriCorps national service program.

Shobita Parthasarathy is Professor of Public Policy and Women's and Gender Studies, and co-founder and director of the Science, Technology, and Public Policy Program at the University of Michigan. Her research examines the political economy of innovation and innovation policy focusing on equity and justice, and the politics of knowledge and expertise in public policy making and law. She often takes a cross-national or international perspective in her research, and has published widely on genetics and biotechnology, intellectual property, innovation policy, inclusive innovation, and artificial intelligence/machine learning in the United States, United Kingdom, Europe, and India. Parthasarathy is the author of numerous scholarly articles and two books: *Building Genetic Medicine: Breast Cancer, Technology, and the Comparative Politics of Health Care* (MIT Press, 2007) and *Patent Politics: Life Forms, Markets, and the Public Interest in the United States and Europe* (University of Chicago Press, 2017). The former influenced the 2013 U.S. Supreme Court case that determined human genes were not patentable; the latter won the 2018 Robert K. Merton Award from the American Sociological Association. She writes frequently for public audiences including in *The New York Times*, *Boston Review*, *Slate*, *Issues in Science and Technology*, *Scientific American*, and *Nature* and co-hosts The Received Wisdom podcast on issues at the intersection of science, technology, policy, and society. Parthasarathy regularly advises policy makers around the world. She is the recipient of fellowships from the Woodrow Wilson International Center for Scholars, American Council of Learned Societies, American Bar Foundation, the Max Planck Institute for Innovation and Competition, and the Wellcome Trust. She holds master's and Ph.D. degrees in science and technology studies from Cornell University, and an undergraduate degree in biology from the University of Chicago.

U.S. District Judge **Patti B. Saris** is a graduate of Radcliffe College '73 and Harvard Law School '76. She clerked for the Massachusetts Supreme Judicial Court and, when Senator Edward M. Kennedy became chairman of the Senate Judiciary Committee, she worked as his staff counsel. She later became an assistant U.S. attorney and eventually served as chief of the Civil Division. In 1986, Saris became a U.S. magistrate judge. In 1989, she was appointed as an Associate Justice of the Massachusetts Superior Court. In 1994, she was appointed to the U.S. District Court for the District of Massachusetts by President Clinton, and she served as chief judge from 2013 to 2019. She served as Chair of the U.S. Sentencing Commission from January 2011 until January 2017. Saris has been active on various judicial committees, including the Budget Committee, and the Defender Services Committee (2002–2005), which she chaired in 2005. Saris currently teaches as a lecturer at Harvard Law School. Saris has received awards for judicial excellence from the Boston Bar Association, the Federal Bar Association, the Patent Bar Association, and the Massachusetts Bar Association. She is a recipient of the Harvard Medal and has received an award for her work in reducing federal sentences for drug offenders.

Appendix: Biographical Information of Committee and Staff

Thomas D. Schroeder is a U.S. District Judge for the Middle District of North Carolina, having taken his oath of office on January 8, 2008. He served as chief judge from 2017 to 2023. Schroeder attended the University of Cincinnati College-Conservatory of Music and received a B.S. in Business Administration from Kansas University in 1981. He received his J.D. from Notre Dame Law School in 1984 and served as editor-in-chief of the Notre Dame Law Review. Following law school, he served as law clerk for the Honorable George E. MacKinnon of the U.S. Court of Appeals for the District of Columbia Circuit. He then practiced law for more than 22 years with Womble Carlyle Sandridge & Rice, PLLC (now Womble Bond Dickinson [US] LLP), where he handled litigation and trials involving a variety of business and product liability matters, many in the capacity as national counsel and frequently involving expert scientific evidence. Schroeder is an elected member of the American Law Institute. He has served as a vice president of the North Carolina Bar Association (2014–15) and co-chair of its bench/bar liaison committee. From 2017 to 2023, he served on the Federal Rules of Evidence Advisory Committee of the Judicial Conference of the United States and chaired the subcommittee that resulted in the 2023 amendments to Rule 702. He has periodically co-taught a class in pretrial litigation as a senior lecturer at Duke University School of Law and is a member of the boards of directors of the Notre Dame Law Association (president, 2021–23) and the Chief Justice Joseph Branch Inn of Court (president, 2016–18).

David S. Tatel served on the U.S. Court of Appeals for the District of Columbia Circuit from 1994 to 2024, succeeding future U.S. Supreme Court Justice Ruth Bader Ginsburg. After graduating from the University of Michigan and the University of Chicago Law School, he served as the founding director of the Chicago Lawyers' Committee for Civil Rights Under Law and then director of the National Lawyers' Committee. He headed the Office for Civil Rights of the U.S. Department of Health, Education, and Welfare during the Carter administration and then founded and led the education practice at Hogan Lovells, where he is now senior counsel. Tatel, along with Dr. David Baltimore, co-chaired the National Academies of Sciences, Engineering, and Medicine's Committee on Science, Technology, and Law from 2016 to 2023. He also chaired the boards of The Spencer Foundation and The Carnegie Foundation for the Advancement of Teaching. Tatel is a member of the American Philosophical Society and the American Academy of Arts & Sciences. He is the author of *Vision: A Memoir of Blindness and Justice* (Little Brown, 2024).

Staff

Anne-Marie Mazza, Ph.D., is the founding director of the National Academies of Sciences, Engineering, and Medicine's Committee on Science, Technology,

and Law. She has served as the study director on more than 20 National Academies reports on topics ranging from science and national security to governance of academic research and emerging technologies to voting technologies to science and the courts. Mazza was the senior staff director for three international summits on human genome editing and has served as senior director of the Committee on Science, Engineering, Medicine, and Public Policy; the National Science, Technology, and Security Roundtable; and the Christine Mirzayan Science and Technology Policy Fellowship Program. She has also served as a senior advisor to the president of the National Academy of Sciences and the president of the National Academy of Medicine. Mazza served in the White House as executive director of the President's Council of Advisors on Science and Technology (2021–2022) and as a senior policy analyst in the White House Office of Science and Technology Policy (1999–2000). She is a fellow of the American Association for the Advancement of Science. She holds a Ph.D., M.A., and B.A. from The George Washington University.

Elizabeth C. Wiggins, J.D., Ph.D., has been on the research staff of the Federal Judicial Center since 1989 and the director of the Center's Research Division since 2020. Among her areas of study are research methods for the study of legal systems, civil case management, bankruptcy law and procedure, judicial workload measurement, court performance standards and assessment, judicial performance feedback, workplace environment surveys, courtroom technology and digital evidence, and access to justice and self-represented litigants. She has helped numerous international judiciaries undertake empirical work to assess the quality of justice and court administration they deliver. She is a fellow of the American Bar Foundation and of the American College of Bankruptcy. She holds a B.S. in psychology and administration of justice from the University of North Carolina at Chapel Hill, a Ph.D. in social and quantitative psychology from Johns Hopkins University, a J.D. from the University of Maryland Francis King Carey School of Law, and an LL.M. from Vermont Law School. Prior to joining the Center, she was on the faculty at Barnard College of Columbia University and a post-doctoral fellow in quantitative psychology at the Ohio State University.

Joe S. Cecil, J.D., Ph.D., is a senior fellow at the Civil Justice Research Initiative at the University of California, Berkeley School of Law. In 2021, Cecil retired from the Research Division at the Federal Judicial Center. While at the Center, Cecil focused on the role of scientific evidence in federal courts, conducting empirical research projects on expert testimony and admissibility of scientific evidence in civil and criminal litigation and the role of court-appointed experts. He developed the first edition of the *Reference Manual on Scientific Evidence* and has worked with teams of scholars to develop subsequent editions. Cecil was a member of the National Academies of Sciences, Engineering, and Medicine's Committee on Science, Technology, and Law, and has served on several National

Appendix: Biographical Information of Committee and Staff

Academies' committees, including study committees that issued consensus reports on eyewitness identification and forensic science. He received his J.D. and a Ph.D. in psychology from Northwestern University.

Steven Kendall, Ph.D., is a senior program officer for the Committee on Science, Technology, and Law at the National Academies of Sciences, Engineering, and Medicine. He has contributed to numerous National Academies' reports, including *Facial Recognition Technology: Current Capabilities, Future Prospects, and Governance* (2024); *Reflecting Sunlight: Recommendations for Solar Geoen지니어링 Research and Research Governance* (2021); *The Emerging Field of Human Neural Organoids, Transplants, and Chimeras: Science, Ethics, and Governance* (2021); *Securing the Vote: Protecting American Democracy* (2018); *Optimizing the Nation's Investment in Academic Research* (2016); *Identifying the Culprit: Assessing Eyewitness Identification* (2014); *Strengthening Forensic Science in the United States: A Path Forward* (2009); and the third edition of the *Reference Manual on Scientific Evidence* (2011). Dr. Kendall has been a staff officer for the National Academies' National Science, Technology, and Security Roundtable since its inception in 2020. Prior to joining the National Academies in 2007, he worked at the Smithsonian American Art Museum and The Huntington in San Marino, California. He received an M.A. in Victorian art and architecture at the University of London and completed a Ph.D. in the Department of the History of Art and Architecture at the University of California, Santa Barbara.

Jason A. Cantone, J.D., Ph.D. is a senior research associate at the Federal Judicial Center. His Center research addresses scientific evidence, civil rules and litigation, federal and state courts, alternative dispute resolution, and court administration/case management. He has also worked with foreign judiciaries on such topics as the rule of law, judicial education, legal psychology, program evaluation, and mediation. Before joining the Center in 2011, he was a U.S. Air Force civilian at U.S. Strategic Command at Offutt Air Force Base in Nebraska. Outside of his Center role, he is an adjunct professor, represented the American Psychology-Law Society on the American Psychological Association (APA) Council of Representatives from 2017 to 2023, and currently serves on the APA Committee for the Advancement of General and Applied Psychology. Dr. Cantone is an APA Fellow and received an APA Presidential Citation for his work related to the criminal justice system. His non-Center research focuses on psychology and legal decision-making and has been published in numerous social science journals and law reviews. He recently published his first co-edited book through Cambridge University Press: *The Cambridge Handbook of Psychology and Legal Decision-Making* (2024).

Meghan A. Dunn, Ph.D., is a senior research associate at the Federal Judicial Center. Her Center research includes survey construction, shared administrative

services, workplace conduct, and the use of courtroom technology and social media by jurors and attorneys. She has co-authored peer reviewed manuscripts on the use of computer animation and juror decision-making and co-authored a social psychology workbook. As part of her work with the Center, she assisted with previous editions of the *Reference Manual on Scientific Evidence*. Dr. Dunn holds a B.A. in history and psychology from Williams College, and an M.S. and Ph.D. in social psychology from Yale University.

Renee Daly was a senior program assistant with the National Academies of Sciences, Engineering, and Medicine's Committee on Science, Technology, and Law (CSTL). Before coming to CSTL, Daly worked with the Board on Animal Health Sciences, Conservation, and Research (formerly known as the Institute for Laboratory Animal Research) in the National Academies' Division of Earth and Life Studies. She received a master's degree in biotechnology from Johns Hopkins University and a bachelor's degree in biomedical science from Virginia Tech.

Dominic LoBuglio was a senior program assistant in the U.S. Science and Innovation Policy Theme at the National Academies of Sciences, Engineering, and Medicine until February 2023. In addition to his work for the policy theme, LoBuglio provided administrative support for the National Academies of Sciences, Engineering, and Medicine's Committee on Science, Technology, and Law. Prior to joining the National Academies in 2019, LoBuglio developed outreach and science education programs within the fundraising department of the Natural History Museum of Los Angeles County. Previously, he also worked in the science and technology division at the Los Angeles office of the Japan External Trade Organization, a section of the Japanese government's Ministry of Economy, Trade, and Industry. LoBuglio holds a B.A. in Japanese language and culture from the University of California, Santa Barbara.